

Beyond Cassini

The triumph of the spacecraft's arrival in orbit around Saturn last week heralds four years of outstanding research. But the longer-term ambitions of NASA and planetary researchers signal trouble ahead.

Few fields of science require more patience than planetary exploration. The scientists who built instruments for the Cassini-Huygens spacecraft now orbiting Saturn began planning for the mission in the 1980s. They worked feverishly to meet a 1997 launch date, then waited seven long years for the first results.

Early signs suggest that the pictures and data now streaming back from Saturn will be worth the wait. Cassini's powerful instruments have already shown new detail in the planet's ring system and revealed the icy satellite Phoebe to be a remnant from the first days of the Solar System. The large moon Titan is proving more mysterious than ever, strangely devoid of the large oceans that many suspected would lie underneath its clouds. For planetary scientists, it looks to be an exciting and surprising four years.

Several of the Cassini investigators are veterans of pioneering missions in the 1980s and 1990s, and can rightly be called the first generation of planetary explorers. What's the outlook for the next generation? Mostly sunny, but with clouds gathering on the horizon.

In recent years, NASA has created a balanced, healthy planetary-exploration programme, at least on paper. Ever more sophisticated Mars probes launch every two years and are producing results. The Discovery programme — modest-sized missions proposed by the research community — has been a scientific and budgetary success. The next spacecraft in the series, MESSENGER, will lift off this month to study the Solar System's innermost planet, Mercury.

As for the outer Solar System, New Horizons will launch in 2006 to the last unexplored planet, Pluto. And NASA is in the early stages of planning its Jupiter Icy Moons Orbiter (JIMO), which will use a nuclear-powered rocket to orbit several of the giant planet's satellites in the course of one mission — impossible using conventional rockets.

All this would be wonderful, but here's where the storm clouds

gather. JIMO is likely to far exceed Cassini's \$3.2-billion cost. Its nuclear power plant has yet to be built. It relies on a heavy-lifter launch vehicle that doesn't yet exist. Nuclear power offers shorter trip times to the outer planets and enough onboard electricity to power instruments far more advanced than any previously sent into deep space, but any new technology is likely to suffer cost overruns and delays. The expected launch date has already slipped from 2011 to 2015.

Add to this uncertainty NASA's plan to send astronauts to the Moon and Mars, a shift in direction that — if it survives this year's US presidential election — will affect every other programme at the agency. Nuclear rockets would be useful for human expeditions as well, and optimists say that having two customers makes the programme stronger than if JIMO were its only purpose. Pessimists worry that this expensive planetary mission will be easy to cut if NASA runs into money trouble with its Moon-Mars plans.

International cooperation is another unknown. Cassini is the historic high point of collaboration between European and US planetary scientists. The Huygens probe was built in Europe, which also footed nearly a quarter of the overall project's bill. But such collaboration seems to be on the wane. The Mars programme, once envisaged as truly international, ended up being mostly American. The JIMO mission is, at this point, NASA-only. If European and US space scientists continue to go it alone, exploration of the Solar System will be more expensive for each side.

After the successful Voyager fly-bys of the 1980s, planetary scientists entered the doldrums, with no new science results for nearly a decade. NASA has a plan to avoid such a hiatus in the future. But it depends on following through with key projects such as JIMO, whose political fortunes bear watching even as we sit back and enjoy the spectacular Cassini images of Saturn, its rings and moons. ■

News cornucopia

How to access all of Nature publications' journalism in one fell swoop.

Now more than ever, people need good news coverage about science. Most media organizations are cutting back on budgets for journalists, including science journalists, who are becoming less and less able to spend time digging into stories, or even searching for them in the first place. Lots of money is being spent on new platforms for disseminating journalism, as the use of news feeds grows and as we head towards electronic replacements for newspapers. But quality, diversity and depth are increasingly in short supply.

This week, *Nature* and its sister journals strike a blow against this trend. Simultaneous with the publication of this week's issue of *Nature*, we are pleased to launch our new online news service, news@nature.com. It delivers free daily stories — a revamped version of *Nature Science Update*, which it replaces. It also provides all the news and other journalistic content provided by *Nature*, *Nature Medicine*, *Nature Biotechnology*, *BioEntrepreneur* and *Nature Reviews Drug Discovery*, as well as career news from *Naturejobs*.

Subscribers to any of the participating journals get all of this online content at no extra charge and before it is printed. Others can subscribe to the news service directly, receiving the free daily service and the rest as premium content. But for an introductory season over the next two months, the entire service is free of charge for all. It can be found at www.nature.com/news.

Quality? We intend it to be of the highest, in keeping with the values of the *Nature* name. Diversity? Our site will bring you hard-hitting daily coverage of all the big developments in research, as well as the science and technology issues that increasingly affect our lives. But there's room too for the scientific fun that lies buried in specialist journals and meetings, and for a more personal view from our team in the form of blogs and columns. And depth? All the journals provide in-depth coverage of their territories issue after issue.

The team that brings you all this is unmatched in its geographical coverage and award-winning talent. We are confident that your supply of the news of science will be greatly enriched by their efforts. ■





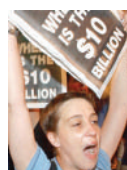
Strike force

The search for dark energy heads for the laboratory
p126



Cut down

Illegal gold-miners disrupt Amazon research
p127



Low show

International AIDS meeting expects fewer US visitors
p128



Off target

Europe struggles to curb carbon dioxide emissions
p130

War of words escalates in run-up to California's vote on stem cells

Jonathan Knight, San Francisco

California is gearing up for an epic battle over a state-wide referendum on whether to issue a \$3-billion bond to support embryonic stem-cell research.

Proposition 71, which will appear on the state's ballot in November, would authorize the money to be spent over ten years and change the state's constitution to protect researchers' right to do the work.

The issue looks set to trigger an almighty row in the state. Many biologists believe that research with stem cells derived from human embryos could yield therapies for degenerative ailments such as Parkinson's disease and diabetes. But restrictions set by President George Bush prohibit federal funding for research on embryonic stem-cell lines derived after 9 August 2001, on the grounds that making new lines involves destroying human embryos.

Proposition 71 is supported by a coalition of patient-advocacy groups, researchers at the University of California's campuses and Stanford University in Palo Alto, and venture capitalists. It would establish a California Institute for Regenerative Medicine to distribute grant money through a review process, prioritizing research not supported by the federal government. By amending the constitution and requiring a 70% legislative majority to make changes to the proposition, the initiative seeks to make California a haven for the research.

"This really separates the funding from the swinging pendulum of politics," says proponent Lawrence Goldstein, a neurobiologist at the University of California, San Diego. "At some point you have to stop arguing and move on."

Opposition to the proposition has brought together people who are against the research on moral grounds and fiscal conservatives — a powerful group in California, who see the bond as something the debt-laden state cannot afford. They have



California could choose to commit \$3 billion to embryonic stem-cell research.

formed a group called Doctors, Patients and Taxpayers for Fiscal Responsibility (DPTFR) and have hired a political consultant to run their campaign.

The measure's cost will probably be one of the most potent arguments against it. Repaying the bond will cost some \$6 billion over 30 years, according to an official state estimate that will appear, by law, on the ballot form.

"This is looting the taxpayers for a very narrow and unproductive area of research that venture capitalists don't want to touch because the benefits are decades away," says Rex Greene, a cancer physician in San Mateo county and a spokesman for DPTFR. Greene says that he is a liberal who supports choice on abortion, but finds himself allied with pro-life Catholics on the ballot issue.

Battle of the bonds

Proponents claim that the initiative will more than pay for itself. Laurence Baker, a health economist at Stanford University who analysed the bond for the initiative, says that spending on buildings and research staff alone will generate enough tax revenue to more than cover the first five years' interest on the bond. Although the long-term benefits are hard to estimate, Baker says that the bond should stimulate the state's

biotechnology industry and give better therapies that will save tens of millions of dollars in medical costs each year.

But with these benefits far in the future, Greene and other opponents say they hope that the public will focus on the measure's short-term beneficiaries. The true winners, they say, are the initiative's sponsors. Most of the \$2.6 million so far raised to support the ballot drive has come from venture capitalists and biotechnology entrepreneurs who stand to gain from its passage, Greene says.

He also questions the new institute's proposed constitution. This involves an oversight committee including members

of patient-advocacy groups, research institutes and five University of California campuses. "The mission isn't curing disease," Greene says. "It's biotechnology entrepreneurship and the aggrandisement of academic research centres."

Fiona Hutton, a spokeswoman for the initiative, says that the measure's checks and balances will ensure money is distributed fairly. She also says the measure emanates from ordinary Californians. "We have a lot of volunteers doing things like mailings and breast cancer walks," she says, although she declines to estimate how many.

California has a history of high-profile, contentious ballot initiatives. The real dark horse in this one is the state's popular and moderate Republican governor, Arnold Schwarzenegger. On the one hand, other Republicans have supported stem-cell research, including former First Lady Nancy Reagan. On the other, Schwarzenegger wants to balance California's budget, and he may hesitate to back a measure whose passage would be seen as a slap in the face for President Bush. His support or antipathy would have a big impact on voters who elected him to fix California's financial woes — but observers say that the former muscleman may choose to sit this fight out. ■

R. HOLMES/CORBIS

Superbug genome excels at passing on drug resistance

Jim Giles, London

A bacterium that is spreading through hospitals because of its resistance to certain antibiotics is well equipped to develop yet more self-protection, according to a study of its genome.

Methicillin-resistant *Staphylococcus aureus* (MRSA) kills about 800 patients a year in Britain alone, and health officials have been watching its spread with alarm (see *Nature* 422, 791; 2003).

A genome sequence for one of its strains, published by a group at the Wellcome Trust Sanger Institute near Cambridge, UK, is significantly different from the six other resistant and non-resistant strains that have been sequenced (M. T. G. Holden *et al.* *Proc. Natl Acad. Sci. USA* 101, 9786–9791; 2004). And the areas that differ contain the genes that confer resistance to drugs.

The finding suggests that resistance genes can transfer very quickly between different strains, says Julian Parkhill, a geneticist in the Sanger team. "This is extremely bad news," he adds. "Strains don't have to develop resistance independently."

Other researchers now fear that many strains of the bacteria could acquire resistance to the antibiotic vancomycin, which is commonly used to treat patients infected with MRSA. Few other effective treatments are available.

In May, a study led by Mark Enright, a microbiologist at the University of Bath, UK, suggested that resistance to vancomycin was developing independently in several strains of MRSA (R. A. Howe *et al.* *Emerg. Infect. Dis.* 10, 855–857; 2004).

Vancomycin-resistant strains were first identified in 1997 and had been thought rare. But Enright looked at about 100 samples of MRSA that had shown some vancomycin resistance. He found that the microbes were related to the five major MRSA strains that cause problems in hospitals, and not to a single strain as some had originally thought.

The implication, says Enright, is that resistance is evolving in strains around the world. "Outbreaks are inevitable as long as we keep using vancomycin," he says.

Researchers do not know when full vancomycin resistance will become widespread. Only three cases have been reported so far, and none led to fatalities. But Enright says that deaths from the bacteria can go unrecorded. "If people are ill anyway, you will not necessarily see MRSA on the death certificate." ■



Give it a whirl: could a simple test detect energy that, so far, has only been measured by cosmologists?

Scepticism greets pitch to detect dark energy in the lab

Philip Ball

A cosmic force that is thought to drive the accelerating expansion of the Universe could be probed using desktop electronics, two researchers have claimed.

The force — usually known as dark energy — seems to oppose gravity, making galaxies fly apart with increasing speed. Detected eight years ago, it presents one of the biggest puzzles in cosmology.

But we may not need high-powered telescopes to study it, according to Christian Beck, a mathematical physicist at Queen Mary, University of London, and mathematical biologist Michael Mackey of McGill University in Québec (C. Beck & M. C. Mackey, preprint at xxx.arxiv.org/abs/astro-ph/0406504; 2004).

They say that noise found in a commonplace electronic device could be the fingerprint of dark energy. If it is, they predict this noise will fall quiet above a well-defined frequency. The idea has yet to be tested.

Other researchers are sceptical, however. "The paper is intriguing," says physicist Peter Milonni of the Los Alamos National Laboratory in New Mexico, "but I think it would be a bit of a stretch to say that it holds water."

Cosmologist Robert Caldwell of Dartmouth College in Hanover, New Hampshire, applauds the idea of lab tests but says that for this one to work, "it would have to violate things we already know to be true".

Dark energy is estimated to account for nearly three-quarters of all the energy in the Universe — but no one really knows what it is or where it comes from. One possible source is a buzzing of empty space known as vacuum fluctuations. Quantum theory implies that a

vacuum is constantly fizzing with particles that pop in and out of existence. Scientists assume that a natural mechanism cancels out the energy this produces, but some wonder if a little survives to give rise to dark energy.

The effects of vacuum fluctuations can show up in the lab. For example, they make two parallel plates separated by a vacuum attract one another, a phenomenon called the Casimir effect. They are also believed to account for some of the noise in Josephson junctions, devices in which an insulating layer is sandwiched between two superconductors.

If vacuum fluctuations are responsible for dark energy, say Beck and Mackey, then the energy density of these fluctuations should be the same as the density of dark energy that has been calculated from astronomical observations.

Calculating the energy density in Josephson-junction noise means adding up the energy at all frequencies. So far, measurements have been made up to almost a trillion hertz. They already give a density that is about equal to cosmological estimates for dark-energy density. So, Beck and Mackey argue, energies at higher frequencies should drop to zero. If that does not happen, they say, we should be looking elsewhere to explain dark energy.

Caldwell says the idea has already been disproved through lab studies of the Casimir effect at very high frequencies, in which no cut-off was observed.

Beck counters that previous estimates of vacuum fluctuations are theoretical, but the Josephson-junction approach offers a direct physical measurement. "The experiment would be cheap and easy to do," he says. ■

Treetop ecologists brought down by miners

Declan Butler, Paris

Tropical ecologist Pierre-Charles Dominique will be out in a helicopter over the Amazon this month, helping to track down illegal gold-miners.

The 'garimpeiros', who have been wreaking damage on the environment for years, have now had a more direct effect on ecological research in the area — last month they ransacked Dominique's research camp, taking material worth €75,000 (US\$92,000) and delaying an ambitious new project.

According to witnesses, illegal gold-miners arrived by canoe at a camp belonging to the Nouragues research station in French Guiana on 12 June. They stripped it bare, taking everything from hydraulic pumps and diesel engines to solar panels and food.

The material was intended for the first trial of the Canopy Operation Permanent Access System (COPAS) — a European project designed to allow researchers easy access to the treetops of a tropical rainforest. When completed, a huge helium balloon will float above the trees suspending a basket that can hold two scientists, who will be able to move vertically and horizontally within the canopy along a system of cables. This will be a boon for researchers who have had to climb trees, swing on aerial walkways or hang from giant cranes to get their data.

The ransacking will delay the project for several months, predicts Dominique. The researchers will also have to ask the French



Dirty tricks: illegal mines damage the Amazon.

National Centre for Scientific Research (CNRS) for extra funds to replace equipment.

Mining has flourished in Guiana over the past two years; there are now at least 10,000 clandestine miners. More than 25 illegal mines are thought to operate in the 1,000 square kilometres of Les Nouragues natural reserve.

These cause substantial ecological damage: river beds are suctioned for gold nuggets, trees are felled, and mud and mercury pollute waterways. Studies show that some 280 tonnes of mercury compounds have accumulated in the environment because of mining in Guiana over the past century, and illegal mining continues to add

5 to 10 tonnes annually. Mercury levels in fish and the local population are also high.

"Without massive intervention by the gendarmerie, it will soon be impossible to pursue the scientific research, carried out since 1986, at the Nouragues station," warned the World Conservation Union following a meeting in Paris on 22 June. On 28 June, the local government in Guiana called on the French president, Jacques Chirac, to establish an emergency plan that "mobilizes all human and material means to have the law respected".

The police have already been given some powers to tackle the situation. In September 2002, they were authorized to destroy on sight all miners' equipment. Special intervention squads have also been created. But search-and-destroy missions are not easy, says Dominique. Illegal sites are readily spotted from the air, he says, but miners deliberately fell trees in nearby clearings to prevent helicopters from landing. This forces the police to make risky rope descents or to approach inland, giving the culprits time to escape.

Several sites have been destroyed in a recent intensification of operations, and the police are also working to dismantle the criminal networks that organize the mining. Dominique says his discussions with local officials have reassured him. "It's too soon to have an official response" from the French government, he says. "But I'm working closely with the local gendarmerie, and they have made Nouragues a priority." ■

'Inspirational' leader quits Madrid heart project

Quirin Schiermeier, Munich, and Monica Salomone, Madrid

A top Spanish cell biologist has spurned the chance to lead a huge cardiology centre under construction in Madrid.

Salvador Moncada, director of the Wolfson Institute for Biomedical Research in London, has been consulting for five years with the National Centre for Cardiovascular Research (CNIC), which is being built by the Spanish health ministry.

Moncada — an expert on the role of nitric oxide in cell signalling — had been planning to take over as the CNIC's scientific director early next year, when about 100 researchers already employed by the centre will move into its new laboratory building.

But immediately after the Spanish elections in March, and just before the newly elected socialist government took power, the health ministry terminated Moncada's consultancy contract. According to sources close to negotiations, officials at the ministry decided that its terms were too generous.

The ministry offered him new terms, but Moncada turned them down and said he would quit the CNIC.

"The rigid Spanish bureaucracy, the excess of authority and, lately, the political defamation have made me take this decision," he wrote in an e-mail to CNIC staff. By defamation, Moncada is believed to mean information leaked to Spanish newspapers about alleged irregularities with business expenses that he claimed while consulting for the centre in 1999. The source of the leak is not known.

Moncada says the salary he has been offered by the ministry is a quarter of what he gets in Britain. "If the government is serious about calling high-profile scientists to Spain it must offer competitive conditions, as Real Madrid did for David Beckham," he says. Beckham is a footballer who left Manchester United for Real Madrid in 2003.

Senior researchers at the CNIC fear that Moncada's departure will spell trouble for the laboratory, which is expected to host up

to 15 internationally recognized groups, including more than 300 scientists, by 2007.

"The huge respect he enjoys in the community was key to attracting some of the most talented young researchers in the field," says Santiago Lamas, a group leader at the CNIC who studies diseases of blood-vessel walls, such as arteriosclerosis.

"It's an immense loss," says Juan Redondo, who leads a group investigating gene expression in blood-vessel walls. "The inspiration and motivation he has given us are not going to be easy to replace."

Redondo hopes that Elena Salgado, health minister in the new government, might yet persuade Moncada to change his mind. But some observers doubt that the left-leaning minister will go out of her way to up the terms of the deal offered by her predecessor. ■

For more news and analysis go to
news @ nature.com
www.nature.com/news

Low US participation clouds AIDS meeting

Erika Check, Washington

The United States has long had a fractious relationship with the biennial international AIDS conferences, and the meeting that opens on 11 July in Bangkok, Thailand, will be no exception.

In 1992, the conference was moved from Boston, Massachusetts, to Amsterdam in the Netherlands because US rules would have prevented participants who were HIV positive from entering the country. The main international AIDS meeting hasn't been held in the United States since.

And this year, the number of US officials and government scientists participating in the meeting has been severely cut. In March,

the Department of Health and Human Services said that it would spend \$500,000 on this year's conference, and send about 50 researchers and officials. By contrast, the agency spent \$3.6 million on the 2002 conference, held in Barcelona, Spain, and sent about 240 people. Health department officials say that the money is better spent fighting AIDS in other ways.

The change is a symbol of the administration's fraught relationship with AIDS activists and researchers. "Spiritually it sends the message that the United States is limiting its participation, and limiting other countries' access to our knowledge and resources," says Judy Auerbach, vice-president for public policy at

the American Foundation for AIDS Research, which lobbies for AIDS research funding.

There have been several other skirmishes between the Bush administration and AIDS activists in the run-up to the conference. The activists acknowledge that the administration has continued to fund AIDS research, as well as starting the President's Emergency Plan for AIDS Relief, an ambitious programme to combat AIDS globally. But they contend that Bush's spending on global AIDS has so far fallen short of the \$15 billion over five years that he promised when he launched the plan in January 2003.

AIDS activists and researchers have also been critical of the Bush administration's attempts to exert direct control over health-related programmes. For instance, two weeks ago, it emerged that the health department has enacted a new policy governing how its scientists interact with the World Health Organization (WHO). If the WHO wants to get scientific advice from scientists at agencies such as the National Institutes of Health or the Centers for Disease Control and Prevention, it must now go through the department rather than contacting the scientists directly.

Critics say that more US scientists and officials should be going to Bangkok to talk about such policies. "There's going to be a lot of discussion about these global issues," says Ben Cheng, deputy director of the Forum for Collaborative HIV Research, a Washington-based group that sponsors meetings on AIDS. "But the groups that are supposed to take a leadership role around HIV are the ones that aren't going to be there."

For more on the AIDS conference, see page 133.



Shouted down: the United States came in for robust criticism during the 2002 AIDS conference.

Biochemist takes the reins at top European laboratory

Quirin Schiermeier, Munich

Europe's flagship molecular biology lab has chosen an insider — Iain Mattaj — to be its fourth director.

Mattaj, a Scot, will succeed Greek biologist Fotis Kafatos as director general of the European Molecular Biology Laboratory (EMBL) based in Heidelberg, Germany, next May. European biologists welcomed the institute's appointment of one of its own researchers as confirmation that the 30-year-old laboratory has found its feet.

"It's a terrific choice," says Michael Ashburner, a geneticist at the University of Cambridge, and former co-director of the European Bioinformatics Institute (EBI) near Cambridge, UK, one of EMBL's four outstations. "It's a tough job, but Mattaj has the skills and credibility it will take."

Mattaj, a biochemist, is one of the most-

cited biologists in his field. He joined EMBL as a group leader 19 years ago and later led the lab's gene-expression programme, becoming its scientific director in 1999. Colleagues describe him as straightforward, humorous and warm, and many say that he is an ideal director for EMBL.

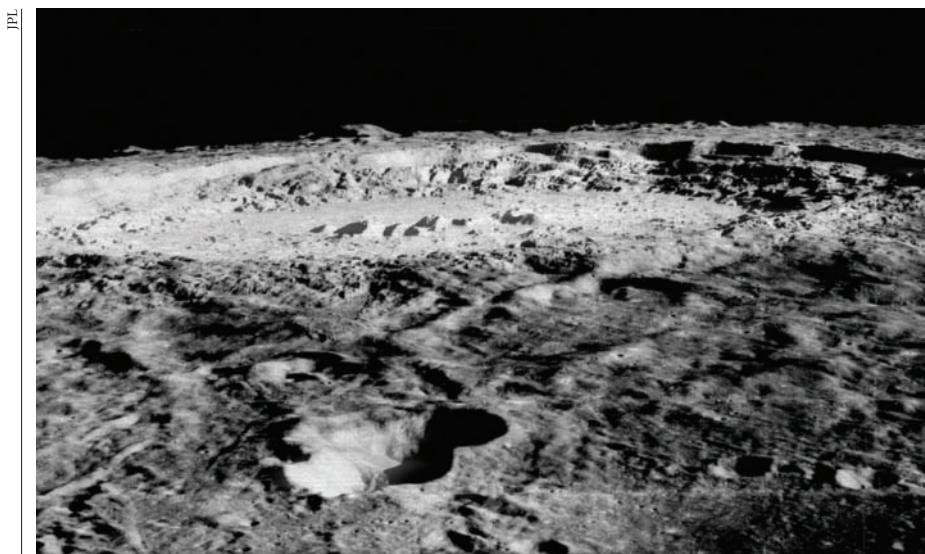
Mattaj says that his priority will be to accelerate a shift in the lab's focus from the molecular biology of individual genes and proteins to systems biology. "Simulation and computer models are penetrating molecular biology with might," he says. "But there's still much good experimental biology that doesn't use computational methods: this tension must be overcome in the next few years."

EMBL is funded by 17 member states, including most of the western European Union, Switzerland and Israel. Its 100 groups, with about 900 scientists, PhD

students and technical staff, work across the spectrum of molecular biology. Besides its headquarters in Heidelberg and the EBI, EMBL operates X-ray facilities for molecular structural biology in Grenoble, France, and Hamburg, Germany, and a mouse-biology programme in Monterotondo near Rome.

The EMBL council last week agreed to postpone by one year the start of the lab's next five-year science programme to give Mattaj time to prepare the right strategy.

He says that he will try to raise extra money for renovation and expansion: the Heidelberg labs need refurbishment, for example. But Mattaj is keen to continue his own cell-biology work. "I sincerely hope I'll still be able to spend some time in the lab," he says. "Otherwise, it is very easy to forget how difficult it is, and how many disappointments are lurking at the bench."



One small step: NASA's push to revisit the Moon's surface starts with a lunar reconnaissance orbiter.

Watchdog slams failings of Israeli animal-rights law

Haim Watzman, Jerusalem

Procedures designed to supervise animal experiments in Israel have not been properly implemented, says the government's main watchdog.

In a strongly worded report released on 29 June, the state comptroller said that processes written into a 1994 law to govern animal experimentation have never come into effect.

The National Regulatory Council on Animal Trials, which was set up in 1994 to establish rules for experiments on animals and to oversee them in universities and other research labs, has not operated in full accordance with the law, the report says.

The animal-welfare law states that scientists should only use animals in experiments if they can demonstrate that no alternatives are available. It says they should restrict the number of animals used to a minimum and never make laboratory animals suffer unnecessarily. "In practice, the issue of alternatives has not advanced much during the council's tenure," the report finds.

The National Committee to Approve Animal Experiments, which the council established to apply the law, has not functioned since its last chairman resigned two years ago, the report says. Since then, individual institutions have established internal review committees to oversee their experiments, as the law permits. But the comptroller says that these committees have not represented non-biologists, as the law requires.

Ehud Ziv, a physician at the Hebrew University of Jerusalem's Medical School who became chair of the council on 1 July, says he welcomes the comptroller's report. He disputes the idea that the council has had no effect, however, citing a continuing decline in the number of animals used in individual experiments. In 2002, some 290,000 animals were used in experiments at 48 institutions in the country, according to figures collected by the council.

Some biologists sought to play down the comptroller's criticisms. "It's a standard report," says Alex Tsafiriri, a biologist at the Weizmann Institute of Science in Rehovot. Tsafiriri is chairman of the Interuniversity Forum for Medical Sciences in Israel, a body set up by Israeli universities to defend the use of animals in science. "It's not surprising to find some creases in the implementation of a ten-year-old law."

Senator urges private groups to run cut-price Moon shot

Tony Reichhardt, Washington

A key Senate committee, frustrated with what it claims are inflated cost estimates at NASA, is calling for the next lunar science mission to be handed to the private sector. But space-agency officials say that letting an outside group run the Lunar Reconnaissance Orbiter mission would be risky and may jeopardize a planned 2008 launch date.

The lunar orbiter is meant to be a first step towards a manned mission to the Moon by 2020, which should in turn pave the way for future trips to Mars. NASA wants the orbiter to create a high-resolution photographic map of the lunar surface, locate mineral deposits and possible water ice from orbit and characterize the threat from radiation to future astronauts.

Even though the Moon is not a new destination, the mission will be challenging, says James Garvin, who heads NASA's lunar science programme. The craft will have to enter into low orbits, carry sophisticated instruments and handle high rates of data flow.

Garvin reckons that the orbiter will cost about as much as a Discovery mission, some \$400 million. NASA estimates that its entire lunar programme, including the orbiter and early work on subsequent lunar landers, will cost some \$1.3 billion between 2005 and 2009. The agency wants the orbiter to be built and operated by its Goddard Space Flight Center in Greenbelt, Maryland, with many of the components contracted out.

Senator Sam Brownback (Republican, Kansas), who chairs the authorizing committee that sets policy guidelines for NASA, calls those price tags "way too high". He says it should be possible to do the first lunar

mission by 2007 for less than \$200 million. The 1994 Clementine lunar mission run by the Pentagon, for example, cost about \$80 million, although much of the technology it used had already been developed.

Brownback has introduced a bill — written in part by adviser Simon 'Pete' Worden, a former US Air Force space official who worked on Clementine — calling for NASA to come up with a new plan for the lunar mission that includes "the use of the private sector to accomplish the goals of the mission".

This could include private companies or non-NASA research groups, such as the Johns Hopkins University Applied Physics Laboratory in Maryland, which has built NASA spacecraft in the past. Splitting the lunar mission into several cheaper spacecraft could be one way of bringing costs down.

Echoing a recent presidential-commission report (see *Nature* 429, 793; 2004), Brownback says "we're not going to get to the Moon and Mars" unless NASA changes its way of conducting space missions. Leaving it to outside engineers and scientists to work out how to accomplish NASA's broad goals should help to kickstart a more diversified space industry, he adds.

But Garvin says that given the tough requirements and tight schedule, the safest course is for Goddard to build this first lunar spacecraft itself. Three years is generally considered barely enough time to mount a space mission, he says.

The debate should be settled over the next few months as the House of Representatives writes its own authorizing bill and the appropriations committees that set NASA's 2005 budget weigh in.

Bioterror charges downgraded to mail fraud for artist

Washington A university geneticist and an artist are facing charges of mail and wire fraud for the way they allegedly obtained bacteria for art exhibitions.

Steven Kurtz, an art professor at the State University of New York in Buffalo, was investigated by prosecutors in early May when laboratory equipment, bacteria and books on biowarfare were found in his home (see *Nature* 429, 690; 2004). The microbes, which he used in performance art, were found to be harmless.

But last week a federal grand jury indicted Robert Ferrell, chairman of the human-genetics department at the University of Pittsburgh in Pennsylvania, accusing him of sending Kurtz some of these microbes illegally. Kurtz could not get bacteria from the supplier as he is not affiliated with a lab.

William Hochul, the chief terrorism prosecutor at the US attorney's office in Buffalo, claims that "the two men went through a whole series of steps" to allegedly defraud the suppliers.

Efrem Grail, an attorney for Ferrell, says his client had no criminal intent. "Dr Ferrell

is a highly respected academic and an ethical researcher who would never do anything he knew to be wrong," Grail says.

If convicted, Ferrell and Kurtz face up to \$250,000 in fines and 20 years in prison.



Europe set to miss Kyoto targets 'by a mile'

London European countries are nowhere near complying with the Kyoto Protocol on climate change, energy economists have warned.

Over the past two months, European Union (EU) states have been working on plans for emission allowances before the

2005 opening of a market for trading carbon dioxide emissions. But almost all the national targets for industrial emissions are above levels needed to meet the Kyoto Protocol, according to the Carbon Trust, a group that lobbies on climate-change issues for the British government. "Member states are going to miss their Kyoto targets by a mile," says Michael Grubb, the trust's director of policy and an energy economist at Imperial College London.

The analysis, published on 2 July, shows that 9 of 12 countries plan to increase emissions in the next three years. But under Kyoto, the EU should cut emissions by 8% from 1990 levels by 2012. If signatories miss domestic targets they must invest in overseas projects to comply with the protocol.

Syngenta joins exodus from UK biotech sector

London The biotechnology company Syngenta is to close its UK laboratories. The Swiss company's decision marks the end of current large-scale commercial research on genetically modified crops in Britain, after withdrawals by fellow biotech giants Monsanto, DuPont and Bayer CropScience.

The decision is "absolutely business-driven", a Syngenta spokesperson told *Nature*. "You have to put your people where



Validated, standardized assays. Now for fluorescence, too.

Introducing the SpectraMax® M2 microplate reader

Molecular Devices' new SpectraMax M2 multi-detection microplate reader with dual-mode cuvette port lets you validate and standardize your absorbance and fluorescence assays in one easy-to-use benchtop unit. And dual monochromators and validation test plates give you the confidence to address your next critical application.

- **Flexible:** choose any wavelength for the widest range of chromophores and fluorophores for your assays
 - absorbance: 200-1000 nm
 - fluorescence: EX 250-850 nm and EM 360-850 nm
- **Powerful:** perform automated instrument validation with SpectraTest™ fluorescence and absorbance calibration plates
- **Smart:** compare different experiments by normalizing data with our patented PathCheck® Sensor
- **Intuitive:** use SoftMax® Pro software for data acquisition and analysis, automation interface and optional GLP/GMP/FDA 21 CFR Part 11 tools

For more information, call 1-800-635-5577 or visit our web site at www.moleculardevices.com/Nature.

 **Molecular Devices**

tel. 800-635-5577 | www.moleculardevices.com

New blood for old remedy as US backs leeches

Paris To the horror of the squeamish but the delight of some medical companies, the United States has decided to approve the marketing of bloodsucking leeches.

As this 1638 medical drawing shows, doctors have used freshwater medicinal leeches (*Hirudo medicinalis*) for centuries, most recently to restore blood circulation to grafted tissues and reattached appendages. Their saliva contains an anaesthetic and an anticoagulating agent, hirudin.

Some companies have sold medicinal leeches in the United States for decades. They can continue to do so, but those wanting to enter the business have to get permission from the



US Food and Drug Administration. On 28 June, French firm Ricarimpex SAS became the first to win approval.

the market is." Most consumers are hostile to agricultural biotechnology in Britain.

The company plans to move its UK crop research to facilities in North Carolina and Switzerland. Syngenta says it plans to relocate staff where possible, but that 130 research and support jobs will be lost.

US bills foreign scholars \$100 for surveillance

Washington Students and scholars wanting to work in the United States will soon have to

pay \$100 each for a system to track their movements.

The new fee, payable from 1 September, will fund the Student Exchange and Visitor Information System, a database designed to keep track of foreigners at US colleges and universities (see *Nature* 427, 190; 2004). The fee will have to be paid before a visa is issued.

Many higher-education groups believe the charge will make it harder for foreigners to work and study in the United States, particularly students from developing

countries without easy access to credit cards or cheques in US dollars. "It's just another barrier to entering," says Victor Johnson, associate executive director for public policy at the Association of International Educators in Washington. "We're concerned that people will start to say 'Why bother?'"

Brazil cuts red tape on research imports

Munich In a bid to strengthen its research base, Brazil has made it easier for scientists to import equipment and materials into the country.

Items cheaper than US\$10,000 will be exempt from import fees and taxes, as long as the purchasing scientists are registered with the Brazilian Council of Scientific and Technological Development. Postal authorities will even fill out researchers' customs paperwork. Such treatment had previously been reserved for non-profit organizations.

The move is being welcomed by researchers burdened with high costs (see *Nature* 428, 453). But many are calling for the government to lift the price limit, according to a news report on *SciDev.Net*, the Science and Development Network website, as much research equipment costs more than US\$10,000.

Innovative Kinase Profiling Tools

Kinase Substrate Sets

Ready-to-screen human phosphosite derived peptides

- ☺ Screen more than 700 potential substrates economically
- ☺ Identification of substrates and consensus motifs

PhosphoSite-Detector

Off-the-shelf peptide micro-arrays

- ☺ Peptide scans through many kinase substrate proteins
- ☺ Identification of phosphosites in substrate proteins

All-in-One Kinase Profiling Service

Full-service using high content peptide micro-arrays

- ☺ Our technology and know-how to profile your kinase
- ☺ Arrays with 25 000 peptides for ultra fast turnaround

Jerini AG Phone: +49-30-9 78 93-117 Fax: +49-30-9 78 93-105 Email: peptide@jerini.com www.jerini.com/peptide

Called to account

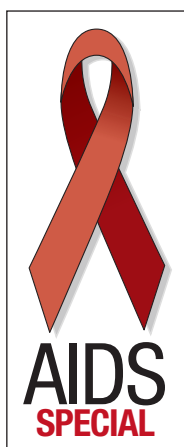
Progress is being made in the fight against HIV, but rich countries have yet to meet all of their pledges to tackle the AIDS pandemic.

Next week, some 15,000 delegates will converge on Bangkok, Thailand, for the XV International AIDS Conference. It is appropriate, given the meeting's location in a fast-developing country that has done much to protect its citizens from HIV, that its theme is 'access for all'. For the poor countries hit hardest by AIDS, this is the crucial issue — they need access to lifesaving drugs, to interventions that can limit the spread of HIV, and to the money to pay for it all.

At the last international conference, in Barcelona in 2002, hopes were high that the rich world would begin to provide the cash to allow developing nations, particularly those in sub-Saharan Africa, to fight back against HIV. "Bangkok will be a time of accountability," observed Peter Piot, executive director of UNAIDS, the Joint United Nations Programme on HIV/AIDS (see *Nature* **418**, 115; 2002). Now it is time to take stock.

Today's balance sheet reveals a mixed picture. On the plus side, more people than ever before are being treated with cocktails of antiretroviral drugs. According to UNAIDS, 230,000 AIDS patients in developing countries were getting these drugs at the end of 2001; two years later, this figure had risen to 400,000. Price reductions have driven this progress — negotiations by philanthropic organizations have helped to lower costs from a minimum of US\$300 per person, per year in 2002 to today's figure of \$140.

But across the developing world, the number of patients requiring antiretroviral drugs outstrips provision by at least a factor of ten. To redress that imbalance, there is an urgent need to cut the cost of drugs that can be used when the virus develops resistance to first-line treatments, and to develop simple and cheap



C. HUGHES/PANOS

blood tests to check whether treatment is actually working. There is also a shortfall in the manufacture of single-pill combined-drug treatments at doses suitable for children.

By 2007, according to UNAIDS estimates, annual global spending on AIDS prevention and treatment will have to rise to \$20 billion per year if the developing world's needs are to be met. Currently, the figure stands at just \$5 billion. At least the trend is positive, with global spending having increased from \$3 billion in 2002. This is partly thanks to the emergence of the Bill and Melinda Gates Foundation as a major player in global health, and the establishment of the Global Fund to Fight AIDS, Tuberculosis and Malaria by the G8 club of rich nations.

The Global Fund, in particular, has been instrumental in convincing poor nations to establish national AIDS treatment programmes. And on 30 June, its board approved grants that should put 932,000 people on antiretroviral treatment. But overall funding is still not assured — currently, pledges from rich nations for 2005 are running at just a quarter of the \$3.3 billion forecast.

In the following pages, *Nature* previews the Bangkok meeting by placing the latest UNAIDS statistics on the HIV pandemic in context (page 134), and profiling a campaigning Thai scientist who is at the forefront of efforts to bring affordable drugs to the African patients who need them most (page 136).

That includes African women who, unable to demand that their sexual partners use condoms, are succumbing disproportionately to HIV. On page 138, we consider whether microbicidal creams or gels could provide them with a degree of protection.

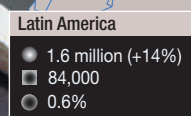
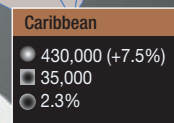
This is just one of the many issues that will keep the halls and corridors buzzing in Bangkok. The long-term dream is still an effective AIDS vaccine. But in its continuing absence, the priority for delegates in Bangkok should be to follow through on Piot's 2002 call to hold the rich world to account. We now have the tools to start bringing the HIV pandemic under control, but this won't be achieved through empty pledges.

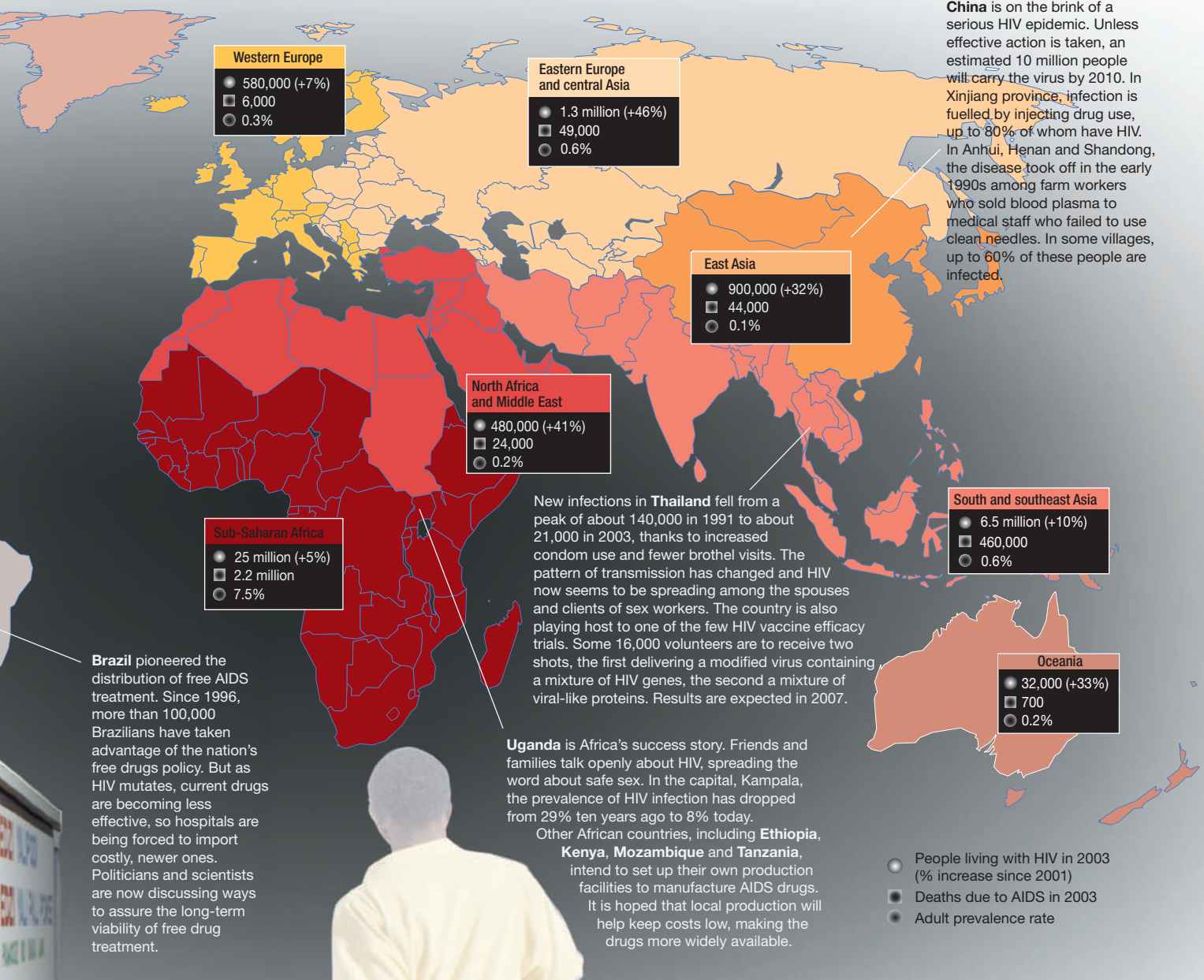
Peter Aldhous, chief news & features editor.
For an online AIDS special visit our new website
news@nature.com at www.nature.com/news

AIDS IS REAL AIDS IS EVERYWHERE

PROTECT YOURSELF AND FAMILY

SPONSORED BY:
NAPCP / WHO AND NATIONAL MILLING & CO.





Snapshot of a pandemic



In 2003, about 4.8 million people became infected with HIV — more than in any previous year. The latest figures from the Joint United Nations Programme on HIV/AIDS, made public on 6 July, show that some 37.8 million people worldwide are now living with the virus.

Those in sub-Saharan Africa, the Caribbean, south Asia and southeast Asia are dying in disproportionate numbers. Last year, the virus killed about 8% of infected people in

these regions, compared with just 1% of those in Western Europe and North America. The high price and scarcity of AIDS drugs in developing countries are to blame.

Sub-Saharan Africa continues to suffer most, with 7 out of every 100 adults carrying the virus. And as Africa's population grows, so too does its problem with HIV. Yet the region has experienced only a 5% increase in the number of people living with HIV between 2001 and 2003, much less than in the rest of the developing world. This is partly because Africa's epidemic is reaching a natural plateau. In some

countries, prevention strategies are also helping to keep the virus in check.

The most alarming surge in HIV infection is in Eastern Europe and central Asia, where there was a 46% rise in the number of people living with HIV between 2001 and 2003. Transmission there is mainly linked to prostitution and injecting drug use, both of which have flourished since the demise of the Soviet bloc. North Africa and the Middle East seem to have experienced a similar increase, but these figures may be unreliable, as surveillance is far from complete.

Helen Pilcher

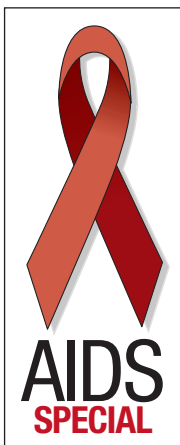
Out of Thailand, into Africa

Can African nations produce their own affordable AIDS drugs? Yes, says the woman behind Thailand's successful HIV treatment programme. Julie Clayton talks to a driven individual who won't take no for an answer.



AIDS warrior: Krisana Kraisintu is determined to bring anti-HIV drugs to everyone who needs them.

Krisana Kraisintu soon realized the magnitude of the task facing her when she made her first trip to the Democratic Republic of Congo (DRC) in December 2002. The instigator of Thailand's successful effort to produce cheap drugs to combat HIV, Kraisintu is attempting a repeat performance where it is needed most: in sub-Saharan Africa.



"I didn't feel any danger until I heard gunfire — there were gunshots and fighting all around us," says Kraisintu. "But I was not afraid to die, because I am a Buddhist, and if I die it is a new beginning."

Despite the continuing civil war, Kraisintu has since returned to the DRC three times. "I will not stop until I've finished," she says. "But it's very slow because of the war." Indeed, a trip planned for March this year had to be postponed because the border with Rwanda was closed.

Kraisintu's goal is to help the DRC, Eritrea and Tanzania to produce generic copies of branded antiretroviral drugs. Only this, she believes, will ensure affordable supplies to support lifelong treatment for impoverished patients. As examples of what can be achieved, Kraisintu points to Brazil and her own

country, which have both established free national AIDS treatment programmes, thanks to local production of generics.

But Brazil and Thailand are among the most industrially advanced nations in the developing world, and many experts don't believe that poor African states are in a position to emulate their achievement. "I don't know how they can, because pharmaceutical production is so technology intensive," says Warren Kaplan, a consultant in public-health and drug policy, currently with the World Health Organization (WHO) in Geneva. "Where are they going to get the people from for the quality control?"

For Kraisintu, such comments echo the scepticism she encountered when she first proposed establishing generic production in Thailand. Her quest was sparked in 1992 by an argument with a prominent Thai politician, who revealed his prejudice against HIV-infected individuals. "He said that all people who contracted AIDS were bad people," Kraisintu recalls. "I quarrelled with him for three hours. I was a member of his political party, so I quit and said that I had to do something."

Fortunately, Kraisintu was in a position to turn her words into action. As director of the Research and Development Institute of Thailand's Government Pharmaceutical Organization (GPO), Kraisintu was well aware of the life-extending and life-enhancing properties of newly emerging cocktails of antiretroviral drugs. But in Thailand they were only within

the reach of the wealthy. Many people with HIV were simply left to die in the care of monks at the nearest Buddhist temple.

Kraisintu began trying to persuade colleagues and government officials that Thailand could produce its own generic versions of antiretrovirals at a fraction of the cost of brand-named originals. "My colleagues didn't want to do it," she says. "The finance department said it would bankrupt the GPO because production would be so expensive."

Winning formula

So Kraisintu had to divert her own precious research funds to purchase the raw materials to produce her first generic AIDS drug, zidovudine, in 1995 — which the GPO sold to the Ministry of Public Health to prevent mother-to-child transmission of HIV.

Over the next four years, Kraisintu developed the formulae for other antiretroviral drugs. Progress was slow, but in 1999 the medical charity Médecins Sans Frontières placed orders for GPO antiretrovirals, attracted by their low prices. This made doctors throughout Thailand realize that the GPO's drugs were safe and effective. Kraisintu was then producing enough drugs to treat only 5,000 patients, still using research funds. But when a Ministry of Finance audit in 2000 showed that the drugs were selling at a profit, the GPO agreed to invest in full-scale production for the treatment of 20,000 patients.

In April 2002, Kraisintu launched a three-



Conflict in Africa is slowing the fight against HIV.

in-one drug cocktail of lamivudine, stavudine and nevirapine, dubbed 'GPOvir'. The single-pill combination made it easier for patients to stick to their treatment. Since then, production has grown even further. By the end of this year, Thai officials plan to expand the national treatment programme to 70,000 people — and are also offering to supply drugs to 30,000 patients in Cambodia, Laos and Vietnam.

These successes are directly attributable to Kraisintu's leadership, says Suwit Wibulpoprasert, deputy secretary at the Thai Ministry of Public Health. "We owe a lot to her." Kraisintu's boundless energy was also a factor. "One week with her is OK, but two or three weeks together is very exhausting," says Christoph Bonsmann of Action Medeor, the German medical charity that is sponsoring Kraisintu's work in Africa.

Kraisintu had first tried to launch a GPO 'Mission to Africa', involving the transfer of technology to make antiretrovirals, in 1999. But her colleagues showed little enthusiasm for the project. Then, in 2002, the German director of the DRC-based company Pharmakina read about the launch of GPOvir in the popular magazine *Der Spiegel*. He invited Kraisintu to the DRC, initially to help produce antiretrovirals for the treatment of infected company employees.

At about the same time, a contact at the WHO recommended Kraisintu to Ramadhan Madabida, director of the drug-manufacturing company Tanzania Pharmaceutical

Industries, who was looking for help in setting up local production of drugs for another devastating disease, malaria. In April this year, with Kraisintu's help, Madabida's firm launched its version of artesunate, an anti-malarial drug derived from a Chinese herb.

Kraisintu decided to make a complete break with the Thai GPO in December last year. In the end, it was an easy decision, as her goals in Thailand had been met. "If I remained in the GPO I would never be able to do anything for African countries," she says.

Now she and Madabida are working to produce cheap antiretrovirals in Tanzania. There is a huge need: more than 2 million of Tanzania's 34 million population are infected with HIV; about 200,000 urgently need drug treatment, yet only some 2,000 are receiving it.

Beating the brands

Kraisintu and Madabida plan to produce the three drugs used in GPOvir. Initially, these will be made as individual pills, and later as a three-in-one combination. Kraisintu expects the locally made generics to cost around 25 times less than the branded drugs on which they are based, allowing each patient to receive combination therapy for as little as US\$140 per year. By September this year, Kraisintu hopes to be producing enough drugs to treat 5,000 Tanzanians, and then to expand over the subsequent few months by between six- and tenfold.

But formidable hurdles lie ahead, including the difficulty of establishing a suitably trained workforce. Kraisintu has set about the task of working in Africa with her customary vigour. On her first visit to Madabida's factory in Arusha, in April 2003, Kraisintu arrived straight from the airport at mid-morning and immediately set about producing antimalarials. "She refused to eat until she got the first result," says Madabida. "By 5 p.m. we had our first tablet. It had a very positive impact."

But getting Tanzanian drug manufacture up to international standards won't be achieved through sheer energy alone. "I'm not sure Krisana is always patient enough," Bonsmann observes.

Kraisintu accepts that she may have to modify the relentlessly energetic approach that underpinned her success in Thailand. "You have to understand the culture," she says. Kraisintu is now learning Swahili, and feels that she is gradually gaining the Tanzanians' confidence — but not without difficulty. For example, she initially faced scepticism when she suggested making artesunate by mixing dry powdered ingredients and compressing them directly into tablets

— avoiding the usual 'wet granulation' step.

Perhaps the biggest obstacle may come from WHO standards for drug safety and quality. Manufacturers can volunteer for WHO inspection, with the incentive that, from 2005 onwards, countries depending on international aid to buy drugs will only be permitted to purchase those that are approved by the WHO.

Tanzania Pharmaceutical Industries' 40-year-old plant will never make the grade — which means that the Tanzanian government will have to fund local production until the facilities can be upgraded. To do so, Kraisintu and Action Medeor have applied for a €5-million (US\$6.1-million) grant from the European Union, to purchase a ready-to-assemble modular production plant designed to internationally acceptable specifications.

The facilities in the DRC and Eritrea are new, and would meet WHO standards, but production of antiretrovirals has not yet begun because of conflict in both countries. Pharmakina has ceased all production and evacuated its staff. But Kraisintu hopes that they will be back by the autumn to start ordering raw materials.

African generic manufacturers may also face legal challenges. In theory, developing countries are protected from patent-infringement suits by amendments agreed to the World Trade Organization's trade-related aspects of intellectual property rights agreement. But in practice, the protective clauses may prove to be unwieldy.

Kraisintu is willing to do battle in the courts if necessary. She was recently preparing to testify against the drugs giant Bristol-Myers

Squibb over a Thai patent for the antiretroviral didanosine (also known as ddI). Kraisintu and others contended that the company had merely licensed the drug from its inventors at the US National Institutes of Health and had made an obvious addition of an agent that enhances absorption into the body.

In January, the case came to a halt when Bristol-Myers Squibb announced that it would "dedicate the patent to the people of Thailand". Kraisintu regrets not having her day in court. "I wanted to fight to the end," she says.

Kraisintu's admirers believe that she will apply the same determination to her work in Africa. "People who are fighting for affordable medicines see her as a warrior," says Onanong Bunjunong, formerly a member of the Thai branch of Médecins Sans Frontières. In squaring up against AIDS in Africa, this warrior has taken on the ultimate fight. ■

Julie Clayton is a science writer based in Bristol, UK.



Starting to gel

In sub-Saharan Africa, there's an urgent need for creams or gels that can protect women from infection with HIV. Now the first large-scale trials are getting under way. Helen Pilcher reports.

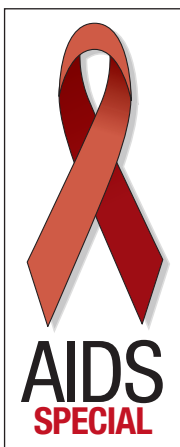
Martha is a fairly typical South African teenager. At 15, she lost her virginity to a man ten years her senior. But he refused to wear a condom, and now Martha is HIV positive.

Martha's is a familiar story in a country where HIV is rife and women bear the brunt of the epidemic. Young women account for nearly 60% of all adult HIV/AIDS cases in sub-Saharan Africa¹. Many find it hard to negotiate safe sex and so cannot protect themselves against the virus. But a new type of drug is being developed to help put women in control.

Microbicides are creams or gels that can be applied to the vagina before sex to help prevent HIV infection. It is hoped that they will be undetectable to male partners, so that women will be able to protect themselves discreetly. With condom use patchy and any successful vaccine some way off, hopes are pinned on these compounds' prophylactic powers.

Women are especially vulnerable in some African cultures, where they are not expected to answer back or to make decisions about their sex lives. It is often hard for them to enforce condom use and difficult to refuse intercourse. Those who do may risk violence. And some impoverished African women have little choice but to exchange sex for material favours, such as shelter, food or school fees.

"I don't think people have an understanding of the enormity of the problem here in Africa," says Ingrid Kelters, an HIV health



worker in Tanzania who runs a tiny health centre on the main road to Zambia, near the stop where long-distance truck drivers break their journeys to buy sex from local barmaids. "You can almost see the HIV flying around at night," she says.

Women are being infected at an early age. In South Africa, one in four women aged 20 to 24 are HIV positive, compared with just one in fourteen men². The epidemic trickles down through the generations as older, sexually experienced men have sex with naive young girls. "This speaks volumes for the need to target therapies at young women," says Helen Rees, executive director of the Reproductive Health Research Unit at the University of the Witwatersrand in Johannesburg.

Shaky start

Microbicides could meet this need. But so far, only one has been fully tested in people, and that proved to be a resounding flop. The detergent molecule nonoxonyl-9 inactivates HIV in laboratory cultures³ and prevents a similar virus, SIV, from infecting macaques⁴. In the United States, it has been sold over the counter as a spermicide since the 1960s, so researchers reasoned it should be safe to use.

But when women used it regularly, it damaged the vaginal lining, an important barrier to HIV infection. More than 400 sex workers from Benin, Côte d'Ivoire, South

Africa and Thailand were enrolled in a clinical trial. Those who used the cream three times per day or more were almost twice as likely to become infected with HIV as those who used a placebo product⁵.

Researchers are now working on gentler microbicides. These include dextrin sulphate and PRO 2000, gels in which large, negatively charged carbohydrate molecules bind to positively charged groups on the surface of HIV, preventing the virus from entering cells lining the vagina. Another microbicide, known as BufferGel, inactivates HIV by lowering vaginal pH.

Six large clinical trials using microbicides have started or are about to begin later this year (see Table, overleaf). The largest, backed by Britain's Medical Research Council and the Department for International Development, will test the ability of dextrin sulphate and PRO 2000 to prevent HIV infection. Over the next few years, 12,000 women in five African countries who are currently free from HIV will apply one of the microbicides or a placebo gel each time before they have sex. They will also receive ongoing HIV counselling and healthcare. Nine months after their treatment begins, the women will be tested to see if they have become HIV positive. "There's a lot of excitement about it," says Rees, who is helping to coordinate the South African arm of the trial.

No one is sure how the gels will perform,

E. HERHOLDT/NDP



although laboratory and animal studies of dextrin sulphate and PRO 2000 look promising. Dextrin sulphate and related compounds stop HIV from entering host cells in culture⁶. And both PRO 2000 and dextrin sulphate provide partial protection from infection by a related virus in female monkeys, who had the gel placed in their vaginas⁷. But it is difficult to extrapolate from monkeys to people — human transmission of HIV is predominantly through sex, whereas monkeys normally transmit SIV by biting and scratching.

Participants in the trial will be encouraged to use both gel and condoms. But it is unlikely that people will use condoms all the time, so the study has been designed to enable statisticians to tease out the effect of the gel alone.

A big difference

At best, microbicide gels are likely to be only partially effective. “But this could make a big difference in places such as Africa,” says Charlotte Watts, who works on the epidemiology of HIV at the London School of Hygiene and Tropical Medicine. Her team has calculated that even a microbicide that was 60% effective could prevent 2.5 million

Bearing the brunt: young women in sub-Saharan Africa face a high risk of becoming HIV positive. Fieldworkers do their best to increase awareness (below) but cash shortfalls hamper progress.

infections worldwide over a three-year period⁸. As well as the obvious health benefits, this could translate into substantial savings for overstretched economies, recouping US\$2.7 billion in healthcare costs and productivity over the same time period.

There could be other benefits too. Condoms provide a high level of HIV protection, but also prevent conception. If gels could be identified that were non-spermicidal, they could offer protection to women who want to conceive but protect themselves from HIV. Microbicides may also help to safeguard against other sexually transmitted infections, such as chlamydia, herpes simplex and syphilis. In turn, this could provide another layer of protection against HIV. Syphilis may cause ulcers inside the vagina, which make it easier for HIV to enter its target cells.

Any HIV-related trial also leaves behind a legacy of improved health education and changed behaviour. Rees's team has recently completed a feasibility study in which 1,000 South African women received healthcare and counselling on HIV and microbicides, but no cream. A year later, condom use had increased.

But hurdles remain, including the need to find microbicide formulations that will be culturally acceptable. In some parts of Africa, such as Zambia, creams may be shunned because vaginal moisture is perceived as a sign of witchcraft. The ideal product would



F. HERHOLDT/MDP

be tasteless, odourless and undetectable to both partners. It should be easy to store and apply, and the applicator — similar to those used for tampons — should be discreet and disposable. Various formulations are in the pipeline. Researchers are testing gels, creams, pessaries and even a silicon-based vaginal ring that could be worn inconspicuously for many months and would slowly release its microbicide. This would free women from the hassle of having to apply a cream or gel each time before sex.

Trials and tribulations

Recruiting thousands of women into microbicide trials is far from trivial. In South Africa, organizers of the British-backed trial are using radio to get the message out. Mthoko Zisi, an HIV/AIDS counsellor for the project, gives talks on local stations to raise awareness. "We've already heard from a range of women who want to get involved," he says. These include rural villagers and city dwellers, some with their partner's support, some without. In turn, these women are being encouraged to talk to their friends in the hope of getting them involved. In Zambia, workers and families on a large sugar estate are being recruited. In Cameroon, prostitutes are being targeted.

At least today, microbicide research is receiving relatively generous funding, after years of being the poor relative of vaccine studies. Over the past six years, the US government has quadrupled its microbicide expenditure and will spend a further \$88 million this year. Worldwide, public- and private-sector spending on microbicides totalled \$530 million over the same period. "But it's not enough," says Polly Harrison, director of the Alliance for Microbicide Development, an advocacy group based in Silver Spring, Maryland. It costs about \$35 million to develop and test a single microbicide. And as new microbicides are developed and more clinical trials are planned, the coffers will need to be topped up.

The next generation of microbicides is already in development. More than 40 compounds are being tested in culture and animal models. It is already known that viral surface proteins, such as gp120, bind to the key receptors on host cells, triggering a wide range of cellular and molecular changes. So researchers are designing strategies that interfere with specific parts of this process. "We know what the targets of HIV are," says Robin Shattock, who works on microbicide development at St George's Hospital Medical School in London. "So we're moving to a

What	Dextrin sulphate and PRO 2000	BufferGel and PRO 2000	Carraguard	Cellulose sulphate	Cellulose sulphate	Saavy
How it works	Both products prevent HIV binding to host cells	BufferGel alters vaginal pH, inactivating HIV	Prevents HIV binding to host cells	Prevents HIV binding to host cells	Prevents HIV binding to host cells	Detergent, destroys HIV directly
Where	South Africa, Zambia, Tanzania, Uganda, Cameroon	India, Malawi, South Africa, Tanzania, Zambia, Zimbabwe, US	South Africa	Benin, Burkina Faso, Kenya, India, Uganda, South Africa	Nigeria	Ghana, Nigeria
Number of participants	12,000	3,100	6,300	2,500	2,200	4,600
Start date	Mid 2004	Late 2004	Late 2004	Late 2004	Late 2004	Early 2004
Expected finish date	2007	2007	2007	2007	2006	2006

phase of rational drug design based on this."

Another possibility is to develop gels containing existing drugs that are used to treat HIV infection. One such drug is tenofovir, which blocks the enzyme reverse transcriptase that is used to copy HIV's genetic material. Researchers hope that a gel containing

tenofovir will last longer than those based on dextrin sulphate or PRO 2000 — it might need applying only once or twice per week, rather than every time before sex.

When a product finally does make it to the market, it will have to be affordable. Dextrin sulphate and PRO 2000 are simple sugar polymers, whose manufacture should be cheap and easy to scale up. The World Bank may subsidize distribution costs, and the US Agency for International Development will help to negotiate low public-sector prices.

Even so, affordability is still a worry, says Rees. The female condom protects against HIV and costs about 70 cents. "But it's not controlling the epidemic because it's too expensive," she says. With 3 billion people living on less than \$2 a day, microbicides will have to be a lot cheaper if they are to be effective.

Fruitful enquiry

Given these concerns, and the fact that the commercial microbicides won't be available until 2010 at the earliest, reproductive biologist Roger Short of the University of Melbourne, Australia, is looking for a cheaper, more readily available alternative. He thinks that the answer may grow on trees.

In a paper to be presented at next week's XV International AIDS Conference in Bangkok, Thailand, Short and his colleagues will report that a 20% solution of lemon juice takes just two minutes to achieve 90% inactivation of HIV in lab culture.

Historically, citrus fruit juices have been used as contraceptives. About 300 years ago, Mediterranean women used lemon juice in

the vagina to help prevent pregnancy. Today, Nigerian prostitutes douche with dilute lemon juice in the hope that it will protect them against sexually transmitted diseases.

Short points out that in South Africa, five lemons can be bought for the price of one condom. And the juice of just one fruit goes a long way. A single lemon yields enough juice for ten sexual acts, and can be kept at room temperature for up to a month.

As long as the vagina is free of lesions, using lemon juice is virtually painless. It could also be applied to the foreskin after sex to help protect men from HIV infection. And because it's natural, it can't be patented. But the flipside is that with no profit motive for drug companies, funding for clinical trials is hard to come by.

Nevertheless, the Thai government recently agreed to launch an initial clinical study. Because limes — *manao* in Thai — grow widely in southeast Asia, the *manao* trial will focus on limes rather than lemons. The study, which should begin shortly, will assess the acceptability, safety and efficacy of lime juice against pregnancy and sexually transmitted infections, including HIV. "We've been looking for the microbicidal Cadillac," claims Short, "when all along we've been overlooking the humble push bike."

Whether the answer lies with Short's low-tech approach, gels such as dextrin sulphate, or advanced rational drug design won't be known until the results from clinical trials are in. But with no sign of an effective AIDS vaccine on the horizon, microbicides may offer the best hope for millions of African young women of avoiding Martha's fate. ■

Helen Pilcher works in Nature's online news team.

- Baeten, J. M. & Overbaugh, J. *Curr. HIV Res.* **1**, 69–86 (2003).
- Pettifor, A. E. *et al.* *HIV and Sexual Behaviour Among Young South Africans: A National Survey of 15–24 year olds* (University of the Witwatersrand, Johannesburg, 2004); http://www.lovelife.org.za/corporate/research/national_survey_RHRU.pdf
- Polsky, B. *et al. Lancet* **i**, 1456 (1988).
- Miller, C. J., Alexander, N. J., Gettie, A., Hendrickx, A. G. & Marx, P. A. *Fertil. Steril.* **57**, 1126–1128 (1992).
- Van Damme, L. *et al. Lancet* **360**, 971–977 (2002).
- McClure, M. O. *et al. AIDS Res. Hum. Retrovir.* **8**, 19–26 (1992).
- Weber, J. *et al. AIDS* **15**, 1563–1568 (2001).
- Microbicide Initiative *The Public Health Benefits of Microbicides in Lower-Income Countries* (Rockefeller Foundation, New York, 2002); www.rockfound.org/documents/488/rep7_publichealth.pdf



Roger Short: believes citrus fruit may offer a cheap and convenient barrier to HIV infection.

Extinction-risk coverage is worth inaccuracies

The media raised awareness of an important issue, even if they got some details wrong.

Sir—Millions of people worldwide learned that climate change poses serious extinction risks to species as a direct result of the news coverage surrounding the Letter to *Nature* by Chris D. Thomas *et al.* (*Nature* 427, 145–148; 2004). Should *Nature* have blocked publicity on this story to prevent possible reporting inaccuracies, as Richard J. Ladle and colleagues (*Nature* 428, 799; 2004) suggest? We don't believe so.

Ladle and his colleagues correctly point out that the time-frame of extinctions was widely misreported. We knew this aspect of the story would be technically difficult, so our press releases in both the United Kingdom and the United States emphasized the correct interpretation in stand-alone paragraphs and italicized key words. The Letter to *Nature* itself emphasized this point, and we stressed the correct

time-frame interpretation to every reporter who contacted us. The majority of reporters to whom we spoke in the United States got the time-frame issue right, although several still misreported it, or fell victim to headline-writers at their news organizations who exaggerated the findings.

Still, the critical issue of connecting species extinctions to climate change was thrust before a broad American public. The story was covered by the most-watched news programme in the United States, the most-listened to radio network, one of the three major news magazines, and five of the top ten newspapers. More than 13 million people saw news programmes on the subject, and total readership of the newspapers covering the story was greater than 21 million. Including radio and magazines, we conservatively estimate that

more than 40 million Americans read, heard or saw a story on this topic. The raised profile of the issue led to testimony being given before the US Senate by one of the study's co-authors.

Breaking through a US media climate often dominated by news of war, terror or the latest celebrity escapades is a victory. We have every obligation to help reporters understand and fact-check their stories before publication, and will continue to commit resources to that effort. But although the reporting wasn't perfect, we believe the benefits of the wide release greatly outweighed the negative effects of errors in reporting.

Lee Hannah, Brad Phillips

Center for Applied Biodiversity Science,
Conservation International, 1919 M Street NW,
Washington, DC 20036, USA

Saving vital time in the war on drug resistance

Sir—Your News story "Defence work sheds light on hospital bacteria" (*Nature* 428, 457; 2004) reported that a British company is developing a luciferase-based test to detect methicillin-resistant *Staphylococcus aureus* (MRSA) directly from patient samples.

According to your News story this new assay could be performed in a few hours, whereas other MRSA tests take days. This is not accurate. For example, a Swiss team has published a six-hour assay combining immunomagnetic enrichment of *S. aureus* cells with real-time polymerase chain reaction (PCR) amplification of species-specific and resistance genes (P. François *et al.* *J. Clin. Microbiol.* 41, 254–260; 2003).

More recently, our group developed a real-time PCR assay that was shown to be sensitive and specific for the detection of MRSA within one hour directly from nasal swabs (A. Huletsky *et al.* *J. Clin. Microbiol.* 42, 1875–1884; 2004). A commercial version of this test, manufactured by Infectio Diagnostic, has now been approved by the regulatory authorities of Canada and the United States and is available on the North American market for the diagnosis of MRSA carriers (www.infectio.com). It will soon be marketed in Europe.

The healthcare burden imposed by resistant bacteria is alarming. The use of clinically validated tests for the rapid diagnosis of MRSA is essential to improve management of antibiotic resistance in hospitals.

Michel G. Bergeron, Ann Huletsky, François J. Picard, Maurice Boissinot

Infectious Diseases Research Center,
Université Laval, CHUQ (pavillon CHUL),
2705 boulevard Laurier, Quebec City RC-709,
Quebec G1V 4G2, Canada

Fat chance of measuring food intake accurately

Sir—The science of dieting, as Declan Butler observes in your News Feature "Slim pickings" (*Nature* 428, 252–254; 2004), is "painfully thin". Yes, but he is too kind. The problem is not merely a shortage of large, long-term, well-controlled trials. More fundamental is our inability to measure what people actually eat.

Most trials of free-living populations involve assigning subjects to alternative diets. Their food intake is measured, if at all, by one of the conventional instruments — food questionnaires, diaries, duplicate portions. All these methods share the same weakness: they depend on people honestly telling researchers what they consumed.

In most trials, there is no independent measure of food intake. 'Independent', in this context, means a biochemical or physiological indicator of the nutrients and other food components consumed. Yet it is universally recognized that most people underreport how much they eat. And overweight people underreport most of all, sometimes by more than 50%. So we do not know to what degree, if at all, subjects were actually following the diets to which they were assigned.

Experiments in sealed metabolic wards have total control over food portions. But these are small, short-term studies in artificial conditions. Getting free-living volunteers to drink double-labelled water — containing markers which allow their urine to be measured for energy expenditure — is an acceptable surrogate measure, but too expensive to be used in the large, long-term trials that are needed.

In contrast, we have precise measures for the dependent variables, such as weight. But these days, serious trials also attempt to measure the effects of alternative diets on other important biochemical indicators, such as serum cholesterol or triglycerides.

With these measures, researchers sometimes draw conclusions that one diet is better for, say, the heart than the other. Yet they do not know if their subjects were actually eating different diets. They are trying to correlate two variables, without having adequately measured one of them.

Only when we can measure accurately both total energy intake and its component parts will we be able to determine which diets yield sustainable weight loss. These measures must be practical for use in trials on large numbers of free-living subjects over long periods. So they must be cheap, rapid, non-intrusive, painless, self-administrable, and capable of direct data transmission. This will require developments in technology as well as in basic science.

In sum, accurate measurement of food intake is the foundation stone for a true science of dieting.

J. T. Winkler

Food & Health Research, 28 St Paul Street,
London N1 7AB, UK

The fight for fair play

Can athletes be held responsible for every substance they take?

John W. Honour

One of the more clandestine games to be played in Athens this August is the competition between the anti-doping laboratories and the drug cheats. Hormones such as anabolic steroids, testosterone, growth hormone and erythropoietin are likely to make the headlines. Yet the source of some of these positive drug tests may be over-the-counter medications and nutritional supplements, rather than banned drugs. This was the case for several tennis players, including UK star Greg Rusedski, who tested positive for the banned steroid nandrolone last year. Rusedski was cleared of any wrong-doing this March after he argued that he had unwittingly taken contaminated electrolyte supplements supplied by the Association of Tennis Professionals (ATP). The new Olympic watchdog, the World Anti-Doping Agency (WADA), wants a firmer line applied to those who test positive for drugs. But as the experience with nandrolone shows, finding the right balance between catching those who knowingly cheat and setting impossible standards is a considerable challenge.

WADA was set up by the International Olympic Committee in 1999 to coordinate and standardize the fight against drugs in sport. At the beginning of this year it finalized an ambitious anti-doping code, to which all sporting bodies must sign up to before this year's games in Athens. The code mandates a list of prohibited substances and specific penalties, including a minimum two-year ban for all athletes testing positive for banned substances. Developing a uniform code from the morass of regulations that govern international sport is a worthy effort. But the reality of drug testing is not always as clear-cut as a positive or negative result would suggest. The problems created by nutritional supplements are a clear example.

Many athletes use nutritional supplements despite advice — from WADA and elsewhere — to avoid them. Supplements can help maintain normal blood-chemistry levels of protein, amino acids, carbohydrates, vitamins, minerals and salts, after training or when travelling. But the scientific evidence for the benefits of supplements is thin, and there are clear risks. An anti-doping laboratory recently tested around 640 non-hormonal nutritional supplements and found low levels of 11 anabolic steroids in 94 of the products¹.

Many competitors at the Olympics will



Free to play: Greg Rusedski was cleared of intentionally taking the banned steroid nandrolone.

be asked to provide urine samples for testing against a range of banned substances that might enhance performance. Doping control was first introduced in 1968 to detect and deter use of around 20 drugs abused in sport. It is now necessary to detect about 150 drugs — and that number increases each year². In 1976, tests for steroids and other hormones were introduced that look for the breakdown products of the drugs (metabolites) as proof that the drug has been in the body, also making it difficult for urine samples to be maliciously 'spiked' to create a false positive. If abuse of a naturally occurring steroid — such as nandrolone — is suspected, then it is necessary to show abnormally high levels of the metabolite. Setting appropriate levels for such substances is more complicated than detecting the presence of a synthetic drug.

As the range of banned substances grows, it is not easy for competitors to be certain that an over-the-counter medication does not contain prohibited substances. Cough medicines containing ephedrine and pseudoephedrine are notorious for causing positive drug tests. Some products now carry warnings against use by anyone who may be

subject to drug testing. But drugs with the same generic name often have different constituents in different countries. There is no global database listing all drugs by trade and generic names, although some national federations provide such information over the phone or Internet³. An athlete needs to be a pharmacist or have such advice available to determine what they can and cannot take.

Nutritional supplements include (or often are) natural products of uncertain and inconsistent composition and at present are not regulated in the same way as pharmaceuticals. Athletes can now buy supplements tested for a limited range of contaminants (although not the 150 on WADA's list), each to a level of 50 parts per billion. But having this information does not necessarily help a player accused of doping. The governing bodies of sport place immense pressure on athletes by a strict liability rule that makes them responsible for everything in their bodies, so ingesting something by accident is no defence.

A better approach would be to consider what level of substance intake is needed to enhance performance, and to revise the rules, including urine thresholds, to

D. PELED/AP

acknowledge this. Research shows that if you take a supplement containing only 1 to 10 micrograms of certain steroids, then a urine sample taken in the next 12 hours may exceed the cut-off level of 2 nanograms per millilitre of nandrolone metabolite needed to fail a drug test⁴. The amount of steroid ingested by mouth in this way has been described as “insufficient steroid to enhance the performance of a gnat”. Compare that with the 200-milligram dose recommended on a daily or weekly basis in the *Underground Steroid Handbook*. Such high doses are injected directly into the muscles, and would register at metabolite levels above 50 nanograms per millilitre for many months. By setting a lower cut-off, the doping laboratories hope to catch cheats that dope themselves between competitions, but a better way of detecting and deterring such activity is by frequent out-of-competition testing, as WADA now requires.

For nandrolone testing, the strict liability rule has led to variable penalties, up to the maximum two-year ban imposed on UK sprinter Linford Christie. WADA claims that the strict liability criteria and sanctions are fair within accepted principles of international law and human rights. But all athletes who tested positive for nandrolone metabolites also tested negatively in the weeks before and after that test. This would not be the temporal profile expected following intramuscular injection. Injected nandrolone stays in the body for months, whereas any supplement is cleared in 48 hours. In cases of suspected testosterone abuse, the rules allow for a period of covert observation before a doping offence is announced. Doping control for nandrolone should follow a similar process. Existing rules are based on thresholds of metabolites in urine, but cannot distinguish systematic abuse from inadvertent intake. All drug tests that use thresholds face this dilemma.

Sporting bodies blame the makers of nutritional supplements, but are not themselves consistent in regulating performance enhancers. Creatine supplements are taken to enhance metabolism at the end of a training session, in the same way as insulin is taken to restore glucose delivery to tissues after exercise or steroids are taken to hasten recovery from injury. Why is insulin, but not creatine, on the WADA banned list? Training at higher altitudes or sleeping in oxygen-deprived rooms has the same effect as doping with erythropoietin or other blood-boosting products, but are tolerated and are probably not detectable in doping tests. These contradictions send mixed messages to athletes.

Governments and industry have tried to improve the quality and labelling of nutri-



Kelli White: stripped of her gold medal from the 2003 World Athletics Championships after admitting taking performance-enhancing drugs.

tional supplements, but legislation is much less rigorous than for pharmaceuticals. New regulations are unlikely to enforce the purity levels that are needed to prevent adverse findings in dope tests. Sportsmen and women will continue to test positive for these compounds, and to fight these rulings in the courts. Under the existing rules and procedures, each case that involves nandrolone needs to be examined individually. WADA needs to look again at the thresholds they set for drug abuse and improve their testing procedures. For example, carbon isotope analysis can distinguish between the metabolites of natural and synthetic testosterone, but this test is expensive and not widely available.

It will be interesting to see if doping scandals in Athens differ from previous Olympic Games in terms of the nature of the drugs and numbers detected. Tests for a new designer steroid — tetrahydrogestrinone (THG) — have been developed in the past two years⁵. The discovery of THG led to bans for UK athlete Dwain Chambers and others who tested positive for the steroid, and is still causing ripples. In the investigations following her positive test for modafinil, US athlete Kelli White admitted to using undetectable steroid and blood-boosting hormones — despite the absence of any positive tests for these substances. The sprinter accepted a two-year ban, excluding her from Athens, and is assisting officials with enquiries. Others who have never failed a drug test are coming under suspicion. Scientists are working hard to develop new tests for muscle-boosting products, such as human growth hormone, or blood-boosting products, such as erythropoietin; if validated in time, these tests might have a major impact in Athens.

Blood tests are being introduced for naturally occurring hormones, such as insulin and erythropoietin, as there are limits to what urine tests can achieve. In cycling, the haematocrit (proportion of red cell volume

of blood) is now measured as a screening test for erythropoietin. Protein analysis of blood samples requires immunological tests, rather than the precise chemical methods used with urine, and have yet to be fully validated. Moving from urine to collecting blood or other tissue samples creates new problems for doping control, not least for athlete acceptance — some athletes will not want to have blood taken.

Some hormone preparations used to provide athletes with banned proteins carry a small risk of prion disease. It seems the determined cheat will try almost anything to enhance performance, regardless of the risks involved. How long will it be before gene therapy is used in sport? Tests are being developed by WADA-

funded research to detect such cheating. WADA is also looking to detect the abuse of ghrelin, a growth-hormone-stimulating protein — a new threat. Doping control will face escalating costs with more tests for the increasing numbers of banned substances — which may be natural or synthetic. New analytical skills will be needed in the anti-doping laboratories.

Even without specific tests for the vogue substances of abuse, supporters of doping control argue that it works as a deterrent to cheating. But others say that it stimulates athletes to find new ways of cheating. And critics say that, over 35 years, there has been no change in the frequency of positive results. With increasing numbers of banned substances and tests, the detection rate might be expected to increase. In addition, most of the positive tests are for the old synthetic drugs rather than the newer natural substances.

Drug tests only work as a deterrent if everyone believes they work reliably and there are no loopholes, which is not the case with nandrolone. Three Austrian rowers who tested positive for nandrolone last August received only a six-month ban from their rowing federation, and so may still be able to attend the Olympic games in Athens. If WADA gets its way, all athletes testing positive in future will be banned for two years, but such strict penalties require unambiguous tests. For nandrolone, they would do well to revisit the rules if they want to catch the real cheats and avoid trampling on the careers of an unfortunate few. ■

John W. Honour is in the Department of Clinical Biochemistry, University College London Hospitals, 60 Whitfield Street, London W1T 4EU, UK.

1. Geyer, H. *et al.* *Int. J. Sports Med.* **25**, 124–129 (2004).
2. World Anti-Doping Agency *List of Prohibited Substances and Methods* online at <http://www.wada-ama.org/en/t3.asp?p=41627&pp=41626>
3. Abbott, A. *Nature* **429**, 689 (2004).
4. Catlin, D. H. *et al.* *JAMA* **284**, 2618–2621 (2000).
5. Knight, J. *Nature* **426**, 114–115 (2003).

An elusive serial killer

The organism behind the three great plagues is still shrouded in mystery.

Plague: The Mysterious Past and Terrifying Future of the World's Most Dangerous Disease

by Wendy Orent

Free Press: 2004. 288 pp. \$25

Return of the Black Death: The World's Greatest Serial Killer

by Susan Scott & Christopher Duncan

Wiley: 2004. 256 pp. £16.99, \$27.95

Richard W. Titball

The threat of bioterrorism has put plague, a disease caused by the bacterium *Yersinia pestis*, into the spotlight. There are two main forms of the disease: bubonic plague develops following the bite from an infected flea and is characterized by the formation of swollen lymph nodes (buboes); pneumonic plague sometimes develops in patients suffering from bubonic plague, and the airborne transmission of the disease to other humans is then a very real possibility. Plague is one of the category A bioterrorism pathogens listed by the US Centers for Disease Control and Prevention in Atlanta, Georgia, so many researchers in the United States have recently redirected their efforts towards understanding the biology of this pathogen. But even before the threat of bioterrorism there was a fascinating story to be told. Books by Wendy Orent and by Susan Scott and Christopher Duncan provide the non-specialist with some insight into this story.

It is generally accepted that there have been three great pandemics of plague. The first, known as the justinian plague, occurred in around AD 550 and was confined mainly to Africa and some parts of the Middle East. The second pandemic probably originated in central Asia and then spread along trading routes into Europe. It is this pandemic, which occurred mainly during the fourteenth and fifteenth centuries but rumbled on into the sixteenth and seventeenth centuries, that is referred to as the Black Death — it is believed to have killed up to 30% of the population of Europe. During the nineteenth century, after initial worldwide spread, the third pandemic of plague was confined mainly to Asia.

Both books do a good job of describing these pandemics. Orent describes all three, whereas Scott and Duncan focus on the Black Death in England — the outbreaks of disease in Penrith and Eyan are especially well researched and documented. Overlaid on these historical accounts are less convincing themes. In an attempt to describe the history of the Soviet bioweapons programme, Orent makes extensive use of interviews with US and Russian scientists. According



Taking its toll: the Black Death was a widespread killer in seventeenth-century London.

to Russian former bioweapons scientists, plague was the favourite bacterial weapon of the Soviet military. Advanced genetic engineering techniques were reportedly used to create strains that were either resistant to antibiotics or that caused unusual forms of disease (although details of the make-up of these strains are not provided). US scientists interviewed by Orent are more sceptical of these claims.

Both books come to the same conclusion but from different angles — that *Y. pestis*, or at least the strains of *Y. pestis* that we are aware of in the West, could not have been responsible for the great pandemics of plague. Scott and Duncan argue that the symptoms and pattern of spread of the disease during the Black Death pandemic are inconsistent with our experience of plague during the twentieth century. They point out that fleas would be inactive during the winter months when cases of plague occurred and suggest that a virus may have been responsible for the Black Death.

Conversely, Orent points out that plague in England during the sixteenth and seventeenth centuries was seasonal, with most cases occurring during the summer months. Scott and Duncan also suggest that the incubation period of the disease was far in excess of that seen with plague. Part of the problem in interpreting the patterns of disease might lie with the assumption that it was either bubonic or pneumonic plague. For example, the authors correctly state that pneumonic plague is not sustainable for more than a few

cycles of transmission. Less correctly, they suggest that pneumonic-plague patients would be too ill to travel and spread the disease. There are actually well documented twentieth-century examples of the disease spreading in this way.

Scott and Duncan also fail to consider some other possibilities. It is generally accepted that a complex interplay between the slowly spreading bubonic form of the disease and the explosive outbreaks of pneumonic plague occurred during the great pandemics, rather than one form alone. Significantly, Orent points out that this exact pattern of disease occurred during the Black Death. Scott and Duncan also fail to take account of the possibility that animal species other than rats played a role in the spread of disease and that fleas and flea bites were much more common between the fourteenth and seventeenth centuries than they are now.

Orent develops a different line of reasoning: that some strains of *Y. pestis*, and especially those found in marmots (large guinea-pig-like rodents) in Asia, are especially virulent. This is an interesting twist because the genome of *Y. pestis* is quite fluid and over past decades some researchers have suggested that hyper-virulent strains of *Y. pestis* might appear periodically.

So what is the evidence that *Y. pestis* was responsible for any of the great pandemics of plague? Well, for the first two pandemics it is not strong enough to refute conclusively any of the arguments presented above. But the symptoms of plague during these great

pandemics match the documented symptoms of today's plague. That *Y. pestis* was the aetiological agent of the third pandemic is irrefutable — the plague bacillus was first isolated and identified by Alexander Yersin during this outbreak. Recent molecular studies with *Y. pestis* provide additional evidence linking the bacterium with the earlier great pandemics. The 'molecular clock' (baseline mutations in housekeeping genes) suggests that *Y. pestis* evolved somewhere between 1,500 and 20,000 years ago, the former figure in remarkable agreement with the appearance of the Justinian plague. Additionally, molecular phylogeny has revealed three genetically defined groups of *Y. pestis*, and these appear to correspond to the groups of strains associated with the three pandemics of plague.

It should be possible to obtain conclusive evidence that *Y. pestis* was the causative agent because a large number of corpses were buried anonymously under towns and cities during the Black Death. Viable bacteria will be long gone, but tell-tale DNA might still be present. But although one research group has reported the isolation of *Y. pestis* DNA from the teeth of presumed plague victims, others have been unable to repeat these studies.

Together these books give a good account of the history of the three great plagues. Other aspects are less convincing. We do not yet have conclusive proof that *Y. pestis* was the cause of the three great plagues. Although I am not yet convinced that the mysterious and unidentified virus suggested by Scott and Duncan was the cause of the Black Death, these books do serve to remind us never to be blinkered to other possibilities. ■

Richard W. Titball is at the Defence Science and Technology Laboratory, Porton Down, Salisbury SP4 0JQ, UK.

..... Clash of the titans

Einstein Defiant: Genius versus Genius in the Quantum Revolution

by Edmund Blair Bolles

Joseph Henry Press: 2004. 256 pp. \$27.95

Arthur Fine

The authors of postmodern novels enjoy the liberty of moving around randomly through time, thereby erasing the appearance of causality and challenging our sense of rational action. In *Einstein Defiant*, Edmund Bolles also zigs and zags randomly through time, but he is not trying to erase the modalities of human understanding. Rather, he is trying to show how entangled that understanding is with the concrete particulars of time, place and culture.

The geniuses of the subtitle are Albert



In dispute: Niels Bohr (left) and Albert Einstein.

Einstein and Niels Bohr, and 'defiant' reflects Einstein's reservations about quantum theory. Although Einstein initially responded with enthusiasm to the new physics, this quickly gave way to disappointment. His reservations were twofold. First, he felt that quantum mechanics had abdicated the historical task of science to furnish us with knowledge of nature that is independent of observers or their acts of observation. The role of the wave function in quantum theory was to provide probabilities for 'results' if appropriate measurements were made (the Born rule). The theory was silent about what, if anything, was likely to be true in the absence of observation. In this sense, it was irrealist. Second, quantum mechanics was essentially statistical and, unlike probabilities in classical statistical mechanics, quantum probabilities were not understood as arising from ignorance of fine details. In this sense, the theory was indeterministic.

Einstein began to probe how strongly quantum theory was tied to irrealism and indeterminism, leading to a series of debates over its interpretation, beginning around the Solvay Conference of 1927 and culminating in the so-called EPR paper of 1935 (A. Einstein, B. Podolsky and N. Rosen *Phys. Rev.* **47**, 777–780). That paper introduced quantum entanglement and laid the groundwork not only for John Bell's investigations into nonlocality, but also for the contemporary development of quantum information theory (computing, cryptography and teleportation).

Einstein Defiant lays out the background of these debates in the context of the 'old' quantum theory, whose duality of waves and particles set the agenda for the new quantum mechanics of 1925–27. It continues the

story up to the Solvay Conference of 1930. In discussions at the conference, Einstein's 'photon-in-a-box' challenged the energy–time uncertainty relation. After a sleepless night, Bohr brought considerations of clock times from general relativity to bear, and saved the day for quantum uncertainty. Bolles uses this triumph as a nice counterpoint to Bohr's earlier defeat over the Bohr–Kramers–Slater theory in which, letting go of strict energy conservation, Bohr had hoped to maintain a wave picture.

By jolting the reader around in space–time, Bolles' nonlinear history produces vivid impressions of times, places and personalities. Moreover, Bolles manages to describe developments in physics with imaginative analogies and contrasts (wave packets as bulges moving through a bull-whip, or the orbit-jumping electron of the Bohr theory as the ghost of Hamlet's father — it's here, it's there, it's nowhere in between). Although equation-avoidance is standard policy in semi-popular science writing, it frequently fails because one well-placed equation may be worth more than a thousand words. Nevertheless, the policy works here because Bolles blends his analogies seamlessly into the historical vignettes around which he weaves his story.

Inevitably, some of the analogies that carry the physics are strained but, overall, Bolles stays true to the physics while keeping the images lively (and his readers awake). One spot that does need attention is his account of the uncertainty formulas, where he confuses the operators p and q with the uncertainties Δp and Δq and overlooks the statistical character of quantum uncertainty.

Of course, it is not only the physics that needs explaining. The philosophical concepts — especially realism and determinism, and Bohr's difficult 'complementarity' — are no more transparent than is the Schrödinger equation. Bolles does an excellent job with probability and determinism — his nuanced discussion of indeterminism in connection with Heisenberg is particularly fine. He fares less well in explaining the issues surrounding realism. Indeed, he makes it appear that Bohr endorsed the idea that measurements actually create the physical attributes that correspond to their outcomes, and that quantum mechanics simply runs counter to the belief that the micro-world of electrons and such is real. As for complementarity, Bolles wisely begs off, noting early on that trying to pin down Bohr is "like trying to cage a beautiful cloud".

All in all, this is a delightful book. It succeeds admirably in capturing the spirit of the people, the times, the science and the intellectual passion that marked the rise of quantum physics. ■

Arthur Fine is in the Department of Philosophy, University of Washington, Seattle, Washington 98195-3550 USA.

Science in culture

Alien visitors

The Science Fiction Museum and Hall of Fame in Seattle opens up a whole new world.

Greg Bear

Perhaps the most fun I've had in Seattle's arts and sciences scene began more than two years ago, with a mysterious e-mail message asking if I would like to advise on a science-fiction exhibition. As a professional science-fiction writer, my response was immediate: of course.

It turned out that Paul G. Allen, investor, philanthropist, co-founder of Microsoft and life-long science-fiction reader, was thinking of creating the world's first comprehensive museum of science fiction.

In the project's exploratory phase, fellow science-fiction author Neal Stephenson and I attended meetings with future curators and museum staff. We set out to determine how such a museum should be designed, and whether it could fit into the space available — sharing a portion of the Experience Music Project building, also developed by Allen. Allen was dead set against just having four stone walls and glass cases gathering dust. Instead he wanted the museum to reflect science fiction's greatest asset: its sense of fun in the face of complicated and even cosmic questions.

That summer, my wife, Astrid, and I escorted a group from the Experience Music Project around the World Science Fiction Convention in San Jose, California, dropping them into the thick of science-fiction fandom — the culture that has nurtured so many science-fiction writers over the decades, including Ray Bradbury and Arthur C. Clarke.

Later we helped a team hired by Allen's Vulcan production company, already convinced of science fiction's impact on scientific and popular culture, to condense the history of science fiction into a handful of succinct themes. Reflecting science fiction's textual origins, the museum designers set out to show that it all begins with the written word: a story, a novel or a screenplay. They also wanted to show the connection between science fiction and science — the symbiotic relationship between imagination and discovery. The challenge became how to prove these two contentions historically and demonstrate them to the public. Over several months, the team put together a comprehensive design featuring interactive exhibits mixed with display cases filled with key artefacts, such as books, magazines, artwork and film props.

The Science Fiction Museum and Hall of Fame, a marvel of passion, technology and design,

opened its doors last month. It represents a proud and dynamic reversal of years of cultural neglect of the story of science fiction — of its history, aesthetic philosophy and symbiotic relationship with science.

The exhibits are organized into five thematic galleries: Homeworld, Them!, Fantastic Voyages, Brave New Worlds and the Science Fiction Hall of Fame. Each gallery reflects a tributary that feeds into the main river of science fiction: these are,

Among the centrepiece exhibits are Captain Kirk's command chair from the original *Star Trek*, a line-up of famous film robots, as well as interactive wall displays showing cities of the future and a wide variety of speculative spacecraft. The Hall of Fame, a curved wall of polycarbonate bricks engraved with glowing portraits, enshrines many of the finest names in science fiction.

One of the most pointed demonstrations of science fiction's impact on technology is the choice



Making an entrance: Seattle's latest tourist attraction invites you to experience the best in science fiction.

respectively, history, aliens and robots, getting from here to there in unusual ways, the changes that will occur in society, and the literary and artistic personalities that created these visions.

Much of the collection is provided by Allen himself, but many display items have been loaned by studios, film-makers, writers and fans. Among my favourites are Poul Anderson's idea notebook, the original manuscript of E. E. Smith's *The Skylark of Space* (the first major novel of interstellar adventure), and many pieces of original art by Chesley Bonestell, including his classic covers from *Collier's* magazine and the lovely *Saturn As Seen From Titan*. Film buffs will meet the original E.T., on loan from Steven Spielberg and Universal Pictures, and the Alien Queen from James Cameron's *Alien* — all terrifying six metres of her.

of Donna Shirley as the museum's director. Shirley was inspired by Robert Heinlein, Bradbury and Clarke to venture into a career in aerospace engineering, and led the team that built the Mars Sojourner microrover, which landed on Mars in 1997. She has loaned a half-scale copy of the rover and a chunk of martian meteorite to the museum.

Science fiction is a product of the poetic, dreaming mind of the sciences. Its considerable influence in books, films and television is ample proof of the need for a well-funded and disciplined study of the field, open to the general public. The Science Fiction Museum and Hall of Fame is an exquisite beginning.

Greg Bear lives in Lynnwood, Washington, USA. His most recent science-fiction novel is *Dead Lines*.

♦ www.sfhomeworld.org

TED S. WARREN/AP PHOTO

Making a difference

Do Animals Think?

by Clive D. L. Wynne
Princeton University Press; 2004. 288 pp.
\$26.95, £17.95

Sara J. Shettleworth

Dogs that learn words the way a child does and crows that make tools catch our interest because their behaviour seems so unexpectedly human-like. Many books on animal mind and behaviour for general readers feed an appetite for seeing animals as being like us, either by detailing the way experiments reveal animals' abilities to count, form concepts and so on, or by describing the behaviour of some particular animals, often primates. Donald Griffin's *Animal Thinking* (Harvard University Press, 1984) or Marc Hauser's *Wild Minds* (Henry Holt, 2000) come to mind as examples of the first genre, and books by Frans de Waal, such as *The Ape and the Sushi Master* (Basic Books, 2001), are examples of the second.

Do Animals Think? most resembles books of the first genre, but with a difference. In this critical account of selected research, Clive Wynne takes aim at over-sentimental anthropomorphism, particularly on the part of animal-rights advocates. He argues that the degree to which animals are like us cannot be the measure of how much they are worthy of our respect and protection. To take one of Wynne's examples, it is fine to take pleasure from watching dolphins leap simply because they look so joyful, but it is important to know the facts about dolphin behaviour and to realize that — contrary to what some sources quoted in the book would have you believe — they have neither self-awareness nor a complex language, nor perhaps take joy in leaping. How can a concern for animals exist alongside an objective understanding that their minds and behaviour are in many respects very unlike ours?

To capture his theme that animals are like us in some ways but not in others, Wynne adopts the not entirely successful simile of a sandwich with two layers of difference



Jumping for joy? The sight of dolphins leaping out of the water can tempt us to be anthropomorphic.

surrounding a layer of similarity. The outer layers are species-specific sensory systems and abilities such as the echolocation of bats, the still mysterious homing of pigeons, and the linguistic abilities of humans. The filling is cross-species similarity in basic instincts and simple learning abilities.

The book is structured like a sandwich, too. Chapters on humans' attitudes to animals and how they can be justified surround a seven-layered 'filling' made up of chapters devoted to the animals themselves, with the sensory, communicative and social systems of honeybees, bats, pigeons and dolphins alternating with chapters on contemporary research into animal minds. This juxtaposition of psychological topics, such as tests of reasoning (from Wolfgang Kohler onwards), tool use, imitation, ape language, mirror self-recognition and theory of mind, together with accounts of the behaviour and life histories of bats, dolphins and bees, is an effective device for underlining the 'otherness' of animal lives and minds.

All this material is presented in a clear, informal and entertaining way, enlivened by historical asides, among which I most enjoyed the speculations on why Kohler was studying chimpanzees on Tenerife in the first place and the passages from an Elizabethan bee-keeping manual. But, at risk of overstraining the comparison, this sandwich contains some substantial chunks of meat that some readers will have trouble swallowing. Wynne emphasizes that misconceptions by non-specialists and overly anthropomorphic interpretations by researchers are rife in areas where animals' behaviour seems most to resemble human behaviour. Moreover, the degree of resemblance between animal and human behaviour is too easily exaggerated, he says. For instance, population-specific

forms of tool use and other learned behaviours in chimpanzees are rightly seen as a form of culture, but it should not be forgotten that even an hour's observation of any group of humans would reveal vastly more complex culturally transmitted behaviours.

Many comparative psychologists who, like Wynne, do experimental research with animals will probably applaud his critical approach to such topics as theory of mind and animal language, and will not find his conclusions controversial. But the final provocative chapter, on why we should care about animals, is unsatisfactory. Wynne rejects first, and in some detail, the idea that animals should have rights. His specific dismissal of the Great Ape Project, which believes that great apes should enjoy the same rights as humans, will raise some hackles. The absence of sound evidence that even chimpanzees have the cognitive capacity for legal responsibility makes the animal-rights position, in his view, as untenable as that of the sixteenth-century villagers who tried rats for eating their crops. The utilitarian view that the recognition of animals' capacity for suffering gives us a guide fares little better, primarily because, Wynne argues, a calculus of suffering is ultimately impossible to apply in detail.

What we're left with, after what is hardly a complete survey of contemporary views, is that it is simply self-evident that we should value and protect animals whether or not they are in any way like us. Perhaps others will take up the challenge of developing a coherent account of how concern for animals is consistent with a dispassionate view of their minds and behaviour.

Sara J. Shettleworth is in the Departments of Psychology and Zoology, University of Toronto, Toronto, Ontario M5S 3G3, Canada.

Reissued classics

The Cosmic Blueprint: New Discoveries in Nature's Creative Ability to Order the Universe

by Paul Davies
Templeton Foundation Press, \$16.95
A return to print for this 1988 classic.

Ben Franklin Stilled The Waves

by Charles Tanford
Oxford University Press, £9.99, \$16.95
A charming historical voyage, ending in the discovery of the lipid bilayer surrounding cells.

Old before their time

Gregory D. Wirth

The discovery of massive, evolved galaxies at much greater distances than expected — and hence at earlier times in the history of the Universe — is a challenge to our understanding of how galaxies form.

Over the past two decades, astrophysicists have been spectacularly successful in explaining the early evolution of the Universe. Existing theories can account well for the time span from the Big Bang nearly 14 billion years ago until the Universe began to cool and form the first large structures less than a million years later. But detailed explanations of how the original stew of elementary particles subsequently coalesced over time, to form the stars and galaxies seen in the present-day Universe, are still being refined. As they report on pages 181 and 184 of this issue, Glazebrook *et al.*¹ and Cimatti *et al.*² have discovered the most distant 'old' galaxies yet. But the existence of these objects at such an early epoch in the history of the Universe seems inconsistent with the favoured theory of how galaxies formed.

That favoured theory is the so-called hierarchical model, in which smaller structures gradually accumulate into ever larger structures, ultimately forming galaxies of the sort we see today³. The most massive galaxies are expected to have formed relatively late in the process, with few existing before the Universe was half its present age. Such predictions can be tested, in principle, through the observations made of distant galaxies.

Nature has provided us with a powerful means of observing the history of the Universe: because the speed of light is finite, as we look out into space we actually peer back in time, seeing distant objects not as they are now, but as they were when their light was emitted millions or billions of years ago. Unfortunately, galaxies more than 6 billion light years away are not only exceedingly faint, but are also particularly difficult to identify. The visible galaxy spectra are 'red-shifted' to longer, near-infrared wavelengths as a consequence of the expansion of the Universe; at these wavelengths, the Earth's atmospheric emission obscures the key spectral 'fingerprints' that are commonly used to identify galaxies.

For these reasons, virtually all of the galaxies known from the early days of the Universe are those that are still forming new stars, and hence emitting copious amounts of light⁴. Although easier to find, such galaxies are not particularly useful for testing theories of galaxy formation because it is impossible to set strong lower limits on how old they are. However, finding significant numbers of



Figure 1 Journey to the early Universe. This image, taken in visible light by the Hubble Space Telescope, shows a plethora of galaxies billions of light years away in a random patch of sky called the Hubble Ultra Deep Field. As part of a survey of galaxies in this region, Cimatti *et al.*² have found several massive galaxies that were already fully assembled billions of years before they should have been, according to current theories. A complementary survey in the northern sky by Glazebrook *et al.*¹ has revealed similar examples of such galaxies, posing a problem for theories of galaxy formation.

massive, evolved galaxies (which finished forming stars long ago) at distances that correspond to half the present age of the Universe would indicate that such galaxies formed much earlier than the leading theory predicts.

Several earlier studies^{5–8} have found evidence for a population of evolved galaxies in the distant past. These studies used the colours of galaxies as rough estimators of their distance — a method that is easier but also much less accurate than measuring galaxy distances directly through the redshift of their spectrum. By pushing some of the largest ground-based telescopes to their limits, Glazebrook *et al.*¹ and Cimatti *et al.*² have now provided the most compelling evidence yet that 'old' (that is, evolved) massive galaxies were numerous at early epochs. Using the 8-metre telescopes at the European Southern Observatory in Chile to survey a small region of the sky in detail, Cimatti *et al.*² report the discovery of four massive, evolved galaxies, all of which are considerably more distant than the previous record holder. Supplementary images from the Hubble Space Telescope (Fig. 1) show that these galaxies appear to be massive and old, both structurally and spectroscopically.

Complementing this discovery is the ambitious Gemini 'Deep Deep Survey', which

was performed using the Gemini observatory's Hawaii-based 8-metre telescope. This study¹ is notable both for the exceptionally long exposures obtained (30 hours for each target) and for the use of an innovative observing mode, which reduces background noise exceptionally well. As a result, Glazebrook *et al.*¹ were able to measure redshifts for far fainter galaxies than is possible by conventional means. As well as strengthening the evidence that massive, evolved galaxies were a significant component of the young Universe, the Gemini team has estimated the change in the abundance of such objects since that time, by observing additional galaxies at intervening distances. They conclude that, going from the present day back to the earliest epochs they probe, the abundance of massive galaxies decreases much more slowly than predicted by the hierarchical model.

With this first solid confirmation^{1,2} that as far back as 10 billion years ago there were already many old massive galaxies, it is clear that even the best models can't fully explain the evolution of galaxies. These studies are forcing astronomers to consider whether massive galaxies grew much earlier than predicted by the hierarchical model, or whether the stars in these earliest galaxies formed in a

NASA/ESA/S. BECKWITH (STSC)/HUDF TEAM

substantially different way from our expectations⁹. As well as providing the motivation to explore new models of galaxy evolution, this is a tantalizing first look at the type of science that will become routinely possible as the next generation of even larger telescopes come online in ten years' time. ■

Gregory D. Wirth is at the W. M. Keck Observatory, 65–1120 Mamalahoa Highway, Kamuela, Hawaii 96743, USA.

e-mail: wirth@keck.hawaii.edu

1. Glazebrook, K. *et al.* *Nature* **430**, 181–184 (2004).
2. Cimatti, A. *et al.* *Nature* **430**, 184–187 (2004).
3. Blumenthal, G. R. *et al.* *Nature* **311**, 517–525 (1984).
4. Steidel, C. C., Adelberger, K. L., Giallisco, M., Dickinson, M. & Pettini, M. *Astrophys. J.* **519**, 1–17 (1999).
5. Dickinson, M., Papovich, C., Ferguson, H. C. & Budavari, T. *Astrophys. J.* **587**, 25–40 (2003).
6. Fontana, A. *et al.* *Astrophys. J.* **594**, L9–L12 (2003).
7. Drory, N. *et al.* *Astrophys. J.* **562**, L111–L114 (2001).
8. Franx, M. *et al.* *Astrophys. J.* **587**, L79–L82 (2003).
9. Somerville, R. S. in *Proc. ESO/USM/MPE Workshop on Multiwavelength Mapping of Galaxy Formation and Evolution* (eds Bender, R. & Renzini, A.). Preprint at <http://arxiv.org/abs/astro-ph/0401570> (2004).

Immunology

Polarizing a T-cell response

Sophie M. Lehar and Michael J. Bevan

Signals through Notch receptors regulate many developmental decisions. New evidence suggests that this pathway is also involved in dictating the tone of the immune response to infection.

We are plagued by pathogens ranging from small viruses to large multicellular parasites, and have evolved a corresponding array of protective immune mechanisms mediated by different 'effector' cells. To clear infections, the immune system must first recognize the type of pathogen involved and then mount an appropriate response. Papers in *Cell*¹ and in *Immunity*² now show that the Notch signalling pathway, best known for its regulation of cell development, may also determine the type of immune response that occurs.

In this case the effector cells are thymus-derived (T) lymphocytes that express an accessory cell-surface molecule known as CD4. These CD4 T cells can become either Th1 or Th2 effector subsets, as defined by their ability to secrete unique combinations of cytokine signalling molecules³. Th1 cells mediate cellular immunity to viruses or intracellular bacteria by secreting gamma interferon (IFN- γ); Th2 cells promote immunity to multicellular pathogens, such as parasitic nematode worms, their signature cytokine being interleukin-4 (IL-4).

T cells do not recognize pathogens directly but rely on other cells — dendritic cells — as intermediaries⁴. Dendritic cells recognize pathogens through receptors that 'see' common determinants found on pathogens. The best characterized of these receptors are the Toll-like receptors⁵. After recognizing a pathogen, dendritic cells migrate to the lymphoid organs, where they interact with T cells, transmitting information about the type of infection encountered and inducing a T-cell response⁴. The upshot of this interaction can dramatically affect the outcome of infection: inappropriate Th1 or Th2 responses are associated with such serious conditions as autoimmunity, allergy or the inability to clear infections³.

A Th1 response is initiated when dendritic cells are stimulated through their Toll-like receptors; this induces the secretion of interleukin-12 (IL-12), which signals undifferentiated (naive) CD4 T cells to differentiate into the Th1 lineage (Fig. 1). Th2 responses are initiated when dendritic cells encounter multicellular parasites or allergens. But neither the receptors that recognize these 'type-2' antigens nor the dendritic-cell-associated molecules that induce Th2 responses are known.

The key cytokine involved in Th2 differentiation, IL-4, is produced by Th2 cells themselves. This is puzzling. If Th2 cells produce the cytokine required for their own differentiation, how is a Th2 response initiated? One possibility is that Th2 cells arise by default when a strong Th1 stimulus is lacking, but dendritic cells can induce Th2 responses directly, even in the absence of IL-4 signals⁶. So it has been proposed that dendritic cells induce differentiation of Th2 effectors through some as-yet uncharacterized pathway.

Amsen *et al.*¹ and Tanigaki *et al.*² provide evidence that the Notch pathway is the missing link. This pathway is an evolutionarily conserved signalling mechanism that regulates lineage choices in a variety of cell types, including T cells⁷. There are four mammalian Notch receptors and five Notch ligands; the latter fall into two structurally distinct classes, Jagged and Delta. When the Notch receptor binds one of its ligands, the intracellular domain of Notch is cleaved to generate an active form of the receptor. This migrates into the nucleus, where it can induce gene expression by activating the transcription factor RBPJ κ .

Amsen *et al.*¹ show that under different conditions dendritic cells can be induced to express either the Jagged or the Delta class of Notch ligands. Delta is induced on dendritic cells exposed to a Th1-promoting stimulus,

lipopolysaccharide, which is a component of bacterial cell walls. This acts through the conventional Toll-like receptor pathway. In contrast, Jagged is induced under conditions that have previously been shown to induce Th2 responses. The authors present evidence that Notch ligands are involved in polarizing the T-cell response towards producing one or the other type of effector cell. They show that dendritic cells that have been engineered to express either Delta or Jagged on their surface promote induction of Th1 or Th2 responses, respectively. They also show that the promoter/enhancer region of the IL-4 gene contains three binding sites for RBPJ κ , and that Notch can activate gene transcription via these sites.

The notion that Notch signals promote production of Th2 cells is supported by complementary findings from studies of CD4 T-cell responses in mice in which the T cells cannot produce RBPJ κ . In separate analyses, Amsen *et al.*¹ and Tanigaki *et al.*² both find that the balance between Th1 and Th2 differentiation is perturbed in these mice. RBPJ κ -deficient T cells differentiate poorly into IL-4-producing Th2 cells, and preferentially develop into Th1 cells producing IFN- γ . The different subsets of T cells regulate the type of antibody produced during an immune response, and Tanigaki *et al.* report that antibody responses are also skewed in mice lacking RBPJ κ in their T cells. The mice have fewer of the antibodies that are normally associated with Th2 responses, and more Th1-type antibodies. Together, these results suggest that Notch signals can alter the balance of CD4 T-cell differentiation into the Th1 or Th2 lineage.

These findings are exciting, but of course questions still remain. How, for instance, are the different Notch ligands able to transduce distinct signals to naive T cells? If the Notch ligands Delta and Jagged can induce CD4 T cells to adopt opposing fates, and Notch promotes Th2 differentiation by activating gene transcription via RBPJ κ , then only Jagged should activate this RBPJ κ -dependent programme. But it is unclear how this distinction is made: in some settings at least, both classes of ligand are able to activate RBPJ κ -dependent signals⁸. Divergent signals could result from the specificity of Notch ligands for different Notch family members. Or Notch ligands might transduce distinct signals through a single receptor, for example through RBPJ κ -dependent or RBPJ κ -independent pathways. The importance of retaining two classes of Notch ligands is underscored by the fact that simpler organisms, which possess a single Notch receptor, also have two classes of ligands that do not signal equivalently⁹. We eagerly await analyses that will unravel the complexities of Notch signalling in naive CD4 T cells. ■

Sophie M. Lehar and Michael J. Bevan are in the Department of Immunology and Howard Hughes

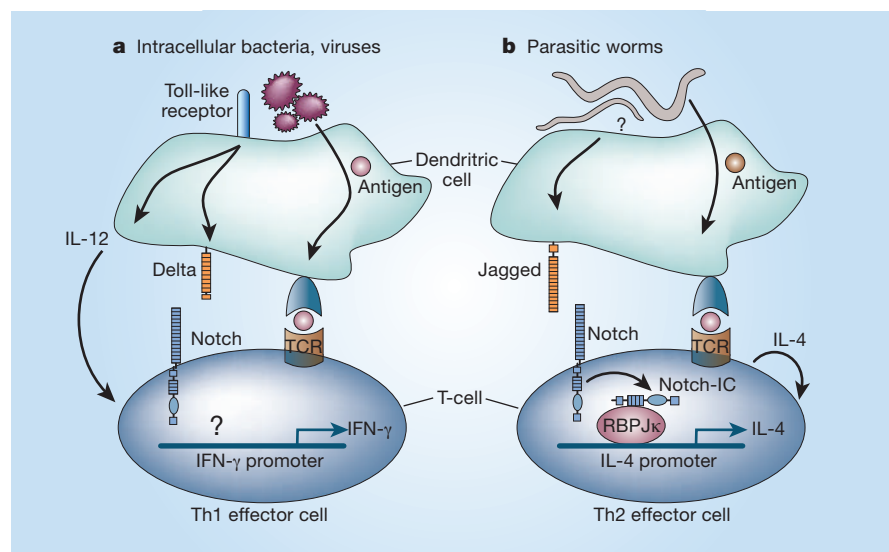


Figure 1 Stimulating the Th1 or Th2 response. In both pathways, dendritic cells internalize the pathogen. They present its antigens to T cells, which recognize antigens through their T-cell receptors (TCR). **a**, Organisms such as intracellular bacteria or viruses are recognized by the Toll-like receptors on dendritic cells; the resulting signals induce the secretion of interleukin-12 (IL-12) and differentiation of CD4 T cells into the Th1 lineage that produces gamma interferon (IFN- γ). **b**, How dendritic cells recognize larger pathogens, such as parasitic worms, is not known. But the end result is differentiation of Th2 effector cells regulated by T-cell-produced interleukin-4 (IL-4). Information^{1,2} on the link between dendritic cells and T cells suggests that the former express different Notch ligands — Delta or Jagged — under different conditions. Jagged is specifically induced by stimuli known to induce Th2 differentiation. Notch signals (Notch-IC) can induce transcription of IL-4 through direct binding of RBPJ κ to the IL-4 promoter¹.

Medical Institute, University of Washington,
Seattle, Washington 98195, USA.
e-mails: smlehar@u.washington.edu
mbevan@u.washington.edu

1. Amsen, D. *et al.* *Cell* **117**, 515–526 (2004).

2. Tanigaki, K. *et al.* *Immunity* **20**, 611–622 (2004).

3. Murphy, K. M. & Reiner, S. L. *Nature Rev. Immunol.* **2**, 933–944 (2002).

4. Kapsenberg, M. L. *Nature Rev. Immunol.* **3**, 984–993 (2003).

5. Barton, G. M. & Medzhitov, R. *Curr. Opin. Immunol.* **14**, 380–383 (2002).

6. Jankovic, D., Sher, A. & Yap, G. *Curr. Opin. Immunol.* **13**, 403–409 (2001).

7. Radke, F. *et al.* *Nature Immunol.* **5**, 247–253 (2004).

8. Jarriault, S. *et al.* *Mol. Cell Biol.* **18**, 7423–7431 (1998).

9. Haimes, N. & Irvine, K. D. *Nature Rev. Mol. Cell Biol.* **4**, 786–797 (2003).

Earth science

Kinks and circuits

Norman H. Sleep

Flow in the Earth's mantle buffets ascending mantle plumes, causing surface 'hotspots' to move relative to each other. A chain of deduction offers solutions to an age-old puzzle about hotspot behaviour.

More than thirty years have passed since the advent of the theory of plate tectonics. Rigid plates and the narrow, deformable boundaries dividing them explain much of the action on the Earth's surface. Plumes of hot material rising from great depth in the mantle are thought to feed 'hotspots', producing surface tracks of volcanism in the middle of plates (such as the Hawaiian islands), and regions of very active mid-ocean-ridge volcanism (such as Iceland). Complicating this picture are zones of deformation, such as the Basin and Range province in the United States, that exist within supposedly rigid plates. In their

paper on page 167 of this issue, Steinberger *et al.*¹ integrate these processes and make them relevant to both field geologists and geodynamicists.

Their work centres on the Hawaii–Emperor bend, a kink in the chain of islands and seamounts that was produced between about 43 million and 50 million years ago, and is evident from even casual observation of a map of the Pacific basin (Fig. 1, overleaf). This feature has long been the 'poster child' example of a change in plate motions; in this case, movement of the Pacific plate is thought to have changed relative to the underlying stationary hotspot. Viewed more



100 YEARS AGO

OXFORD. – The following is the text of [a speech] delivered by Prof. Love in presenting recipients of the degree of D.Sc. *honoris causa* at the Encaenia, June 22, in the presence of the Chancellor of the University:—
THE HON. CHARLES ALGERNON PARSONS.
Duobus fere millibus abhinc annis Heron Alexandrinus turbinem quemdam per ludum excogitavit, qui vapore calido actus per tubos inflexos afflante converteretur. Carolus Algernon Parsons inter Hibernos nobilissimus, scientiae etiam laude insignis, ita Heronis vestigiis institit ut, quod ille ludendi causa finxerat, ipse in usum nostrum converteret, quo facilius homines naturae imperarent. Optime sane meritis est de omnibus qui urbes habitant, quibus vias et domos luce electrica hoc invento usus illustravit, neque minus profuit Nerea temptantibus, cum his turbinibus impulsae per altum naves celeritate inaudita ferantur recta semper carina adeo ut navigantium incommoda magna ex parte adleveretur.
From *Nature* 7 July 1904.

50 YEARS AGO

The main point I wish to make, therefore, is that success in analysing biological molecules by X-rays may explain to us why they have the structure they have. To make a comparison once more with the mineral world, we see that silicon and oxygen together build very stable structures which are light and have a high melting point, and that is why most of the earth's crust is made of silicon and oxygen. Their tetrahedral frameworks conveniently accommodate certain other elements, and correspondingly it is found that these rank next to silicon and oxygen in order of frequency of occurrence in the rocks. On the other hand, carbon, nitrogen and oxygen, together with hydrogen, build structures which are relatively unstable, but which are capable of an infinite complexity. This is so because the atoms are fastened together by bonds which have definite positions. Hence Nature has used these elements to make the complex structures of living matter. Now, as in the case of the silicates, we shall perhaps be able to see further into the reason for the arrangements of these elements which Nature actually uses. If in due course we make a voyage to Mars by a rocket-ship, we can confidently predict that mineralogy will be very much the same on Mars as it is on this globe of ours.
Sir Lawrence Bragg
From *Nature* 10 July 1954.

carefully, however, hotspots move with respect to each other and to the Earth's spin axis. Geodynamicists want to resolve these effects independently of changes in relative plate motions.

Mantle plumes are merely upwelling conduits of hot, low-viscosity material in the mantle flow, and clearly must move with respect to each other: the conduit ascends buoyantly while being dragged rather like smoke from a chimney on a windy day. Steinberger *et al.*¹ model the fluid dynamics of this 'mantle wind' by including the observed plate motions. Much of the conduit height is within the very viscous lower mantle, which flows slowly, giving the illusion (and convenient approximation) that hotspots are fixed. They are not, but most of the time move at less than a centimetre per year. Although the Hawaiian hotspot moved faster at some times, it did not do so in a way that would create a sharp bend in the Hawaii–Emperor track.

Steinberger *et al.* then direct their attention to inadequacies in understanding the relative motions between plates. In case you think this has been sorted out to decimal places in the past 30 years, it hasn't. We can calculate relative plate velocities in the past only where plates were in contact at ridge boundaries (where plates are separated by a ridge, creating new crust) or transform boundaries (where they are sliding past one another). At subduction zones, where one plate slides beneath another, the information about past behaviour, encapsulated in palaeomagnetic data in the sea floor, becomes lost. Mathematically, the relative velocity $\mathbf{v}$ between plates A and B is $\mathbf{v}(A,B) = \mathbf{w}(A,B) \times \mathbf{r}$, where $\mathbf{w}$ is the rotation 'pole' vector and $\mathbf{r}$ is the radial vector to the surface point where velocity is measured.

To study the fixity of hotspots, one needs to estimate the velocity of plates, such as the African and Pacific, that are not in contact by ridge boundaries. Formally, this requires summation of rotation vectors around a 'plate circuit', $\mathbf{w}(A,B) + \mathbf{w}(B,C) = \mathbf{w}(A,C)$. This procedure becomes problematic if, at some point in the circuit, two plates have only a short boundary. Then one cannot accurately determine the component of the rotation vector in the direction of $\mathbf{r}$, as it produces no velocity. This component, however, produces velocity elsewhere when added to the circuit. In connecting up plate relationships around the circuit, Steinberger *et al.*¹ carry through errors in their analysis and keep this problem in check.

Next, the authors had to take into consideration the fact that some plates are not really rigid, especially within continents, which precludes a simple circuit. For times before 43 million years ago (older than the Hawaii–Emperor bend), they get considerably different results by treating the Antarctic plate as rigid, or alternatively taking the circuit

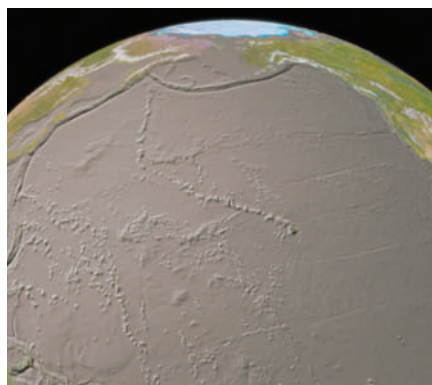


Figure 1 **Spot the bend.** Start at Hawaii and follow the features trending west-northwest until they take a sharp turn north.

through the extinct mid-ocean ridge in the Tasman Sea between Australia and the Lord Howe Rise — a continental fragment that was then attached to New Zealand on the Pacific plate. They prefer the latter solution (see Fig. 3 on page 170) as it gives the bend in the Hawaii–Emperor chain at the right time.

This work makes sufficiently precise predictions to interest the field geologist. For example, the models imply that deformation within the New Zealand plate occurred between 65 million and 83 million years ago, and that there was considerable deformation within Antarctica between 43 million and 83 million years ago. These are surmises that can be checked by field work.

To test their calculations relating to the Hawaiian hotspot, Steinberger *et al.* included analyses of three other hotspot tracks in different parts of the world that have produced island or seamount chains—Réunion (in the Indian Ocean), Louisville (in the Pacific) and Tristan (in the Atlantic). As with Hawaii, an important constraint comes from the estimation of palaeolatitudes — shifting latitudinal position in the past — especially of volcanic edifices that are still submerged. The reason that Steinberger *et al.* worked with only four tracks is that, perhaps surprisingly, numerous submarine edifices have not yet been sampled and dated, let alone drilled

for palaeomagnetic studies. But the four that they have worked with provide consistent results, supporting the conclusion that there was a significant change in relative plate motions at the time that the Hawaii–Emperor bend was created.

What does such a change imply for the past and present geodynamic behaviour around the circuit? Plates move far too slowly for inertia to be relevant. Buoyancy forces associated with plate age in the crust and uppermost mantle, and with slab material in the deep mantle, evolve too slowly to realign plate motions quickly. But factors associated with the shallow parts of subduction could have a considerable effect.

In search of a mechanism for an abrupt plate reorganization, Steinberger *et al.* point out that the Antarctic plate may have become strong enough to be rigid around the time the Hawaii–Emperor bend occurred, and that this may have triggered a change in the circuit, with slab subduction being initiated along a transform fault boundary in the Philippine Sea; today this is the Marianas subduction zone. Once started, subduction provides a driving force. The viscosity of the mantle increases with depth, so much of the driving force comes from the upper few hundred kilometres of the slab. That is, once it was under way, subduction beneath the Marianas quickly increased the change in plate motions.

I expect that debate will continue on the relative fixity of hotspots, the rigidity of tectonic plates and mantle dynamics. Meanwhile, more geological data will come in hand, and seismological studies will continue the task, just begun, of resolving the current behaviour of plume conduits². Such work will batten down our understanding of the present-day effects of the mantle wind. ■

Norman H. Sleep is in the Department of Geophysics, Stanford University, Stanford, California 94305, USA.
e-mail: norm@geo.stanford.edu

1. Steinberger, B., Sutherland, R. & O'Connell, R. J. *Nature* **430**, 167–173 (2004).
2. Montelli, R. *et al. Science* **303**, 338–343 (2004).

Ion channels

Gate expectations

Maria L. Garcia

The opening and closing — gating — of ion channels in response to specific stimuli is crucial for cell function. The membrane-partitioning activities of two venom toxins give insights into the mechanisms involved.

The membrane surrounding a biological cell forms a highly selective barrier that allows the cell to control its internal environment. It consists of an assembly of lipids and proteins. Some membrane proteins form channels through which ions

can flow. These ion channels are said to be 'gated' if they can be opened and closed, and the trigger for this can be electrical, chemical or mechanical stimuli. Ion channels are widely distributed throughout the human body, and in many cells different

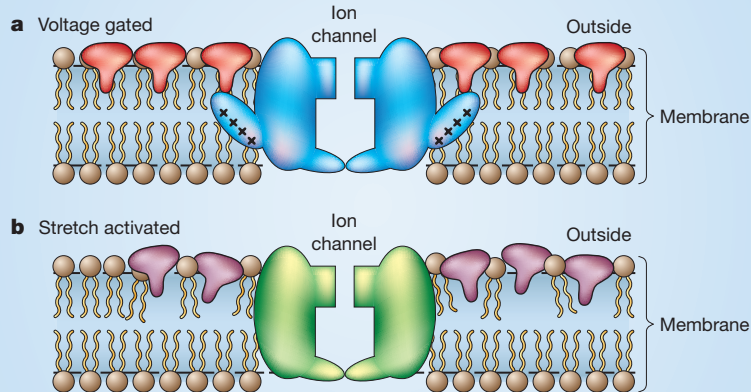


Figure 1 Inhibition of ion channels by two gating-modifier peptides. Schematic representation of a voltage-gated potassium channel (a) and a stretch-activated channel (b). The peptides VSTX1 (a, red) and GsMTx4 (b, purple) partition within the membrane. Lee and MacKinnon² show that the gate in a is favoured in the closed conformation when VSTX1 reaches the positively charged voltage sensor through the membrane, whereas Suchyna *et al.*³ show that opening of the gate in b is blocked through alterations in lipid packing at the interface between the membrane and channel.

types of channel coexist. It is their interplay that tunes the cell's electrical activity, enabling it to function properly; even small changes in the balance of activity of the channel proteins can have significant physiological consequences, sometimes leading to severe medical conditions¹. The development of therapeutic agents that target ion channels is therefore under way.

Now Lee and MacKinnon² and Suchyna *et al.*³ (pages 232 and 235 of this issue) describe how two toxins isolated from the venom of tarantulas interact with two different types of ion channel. Remarkably, they find that the toxins' high-affinity block of these channels is not due to a complementary interaction between channel and toxin surfaces, but to the toxins' ability to partition into the membrane.

The deadly effects of spider venoms are generally thought to result from the high-affinity binding of their toxin components to membrane ion channels. Many toxins are peptides — small assemblies of amino acids — that specifically modulate ion-channel activity. Several peptides that modify the gating of ion channels have been identified. Members of this class interact with the families of voltage-gated channels for sodium, calcium or potassium ions, and with mechanosensitive channels, whose gating can be altered by mechanical forces⁴. Gating-modifier peptides seem to have a common structural feature in that one face is almost exclusively hydrophobic, which seems to be important for affinity⁵. The peptides are thought to access their receptor on the channel from the outer, extracellular, side of the membrane and to bind to regions on the protein that are normally involved in the conformational changes, triggered by a given stimulus, that lead to channel opening⁶.

But gating-modifier peptides have two characteristics requiring special considera-

tion. First, whereas the kinetics of other channel-inhibitory peptides that bind at the outer, surface-exposed region of the channel pore are generally consistent with a diffusion-controlled process, gating-modifier peptides act far more slowly. And second, the interaction surface on the channel for gating-modifier peptides seems to be too small to account for their high affinity.

In an attempt to explain these discrepancies, Lee and MacKinnon² investigated the interaction of a voltage-sensor toxin, VSTX1, with a voltage-dependent potassium channel known as KvAP. The structure of this channel has been determined from high-resolution X-ray crystallography by MacKinnon's laboratory^{7,8}, making it a good study candidate. Suchyna *et al.*³, meanwhile, looked at the venom peptide GsMTx4, which inhibits stretch-activated ion channels.

Lee and MacKinnon found that if the KvAP channel is removed from the membrane using detergent, VSTX1 binds to the solubilized protein with a 10,000-fold lower affinity than when the channel is present in the membrane. Suchyna *et al.* show that GsMTx4 has the same efficacy as its enantiomer — its mirror-image structure — in modifying native stretch-activated channels. They obtained similar results with artificial membranes consisting of lipid bilayers and channels formed by a peptide called gramicidin. These results fly in the face of the traditional 'lock-and-key' model of ligand-receptor binding, and instead suggest that the peptide not only acts on the ion channel but also alters the lipid packing at the interface between the channel and the membrane. Thus, much of the binding energy of VSTX1 and GsMTx4 is derived from the nonspecific free energy of membrane partitioning, whereas the actual peptide-channel interaction is rather weak (Fig. 1).

The high-resolution X-ray structure of

KvAP has led to a controversial proposal for voltage-dependent gating that involves a 'voltage-sensor paddle', a positively charged voltage sensor (Fig. 1a). The paddle model is very different from conventional gating models⁹ in both the shape of the channel in the closed state and the mechanism by which membrane depolarization leads to opening of the ion pathway. Studies of another gating-modifier peptide called hanatoxin have shown that, when this peptide is applied extracellularly, it can bind to closed channels. This has been interpreted as evidence that the voltage sensor lies on the outer part of the membrane in the channel's closed state, thereby limiting its movement within the membrane during activation¹⁰. Furthermore, the slow kinetics of hanatoxin-induced block suggested that it remains bound to the channel during gating. However, these observations must be reconsidered in the light of Lee and MacKinnon's results². If hanatoxin can diffuse into the membrane, previous results would be compatible with a paddle model of gating whereby the voltage sensor moves within the membrane during depolarization. Moreover, membrane partitioning combined with low-affinity binding could allow the peptide to dissociate and rebound rapidly during gating.

How do the findings in these two papers^{2,3} bear on our understanding of therapeutic agents that target ion channels? Traditional pharmacological methods provided the ion-channel drugs currently in use, but drug development is now based on a more rational approach. An ion channel is usually selected as a potential therapeutic target, and possible channel modulators are then screened for activity. Ideally, however, structural information would be used to increase the potency and specificity of lead candidates. Knowledge of the conformations that ion channels assume during gating, as well as the precise nature of drug-channel interactions, will lead to more effective drugs. Because gating-modifier peptides seem to bind to a region of the channel that differs between channel families and even within the same family, it was assumed that drug specificity might be achieved by using a small molecule that mimics the peptide's effect on its target channel. But these two papers suggest that such an approach might not work because nonspecific membrane partitioning contributes greatly to the high affinity of these peptides.

There is one piece of good news, though. The observation that the two enantiomers of GsMTx4 are equally effective in blocking stretch-activated channels has major implications for the use of peptides in drug therapy. This is because, generally, only one of the two enantiomers is susceptible to the body's breakdown enzymes, so the alternative form might be stable in the circulation. Finally, although other small organic

molecules seem to modify gating by accessing the channel through the lipid membrane, it remains to be determined whether membrane partitioning is a common mechanism for all ion-channel gating modifiers.

Maria L. Garcia is in the Department of Ion Channels, Merck Research Laboratories, PO Box 2000, Rahway, New Jersey 07065, USA.
e-mail: maria_garcia@merck.com

1. Ashcroft, F. *Ion Channels and Disease* (Academic, New York, 2000).

2. Lee, S.-Y. & MacKinnon, R. *Nature* **430**, 232–235 (2004).
3. Suchyna, T. M. *et al.* *Nature* **430**, 235–240 (2004).
4. Ruta, V., Jiang, Y., Lee, A., Chen, J. & MacKinnon, R. *Nature* **422**, 180–185 (2003).
5. Wang, J. M. *et al.* *J. Gen. Physiol.* **123**, 455–467 (2004).
6. Winterfield, J. R. & Swartz, K. J. *J. Gen. Physiol.* **116**, 637–644 (2000).
7. Jiang, Y. *et al.* *Nature* **423**, 33–41 (2003).
8. Jiang, Y. *et al.* *Nature* **423**, 42–48 (2003).
9. Ahern, C. A. & Horn, R. *Trends Neurosci.* **27**, 303–307 (2004).
10. Lee, H. C., Wang, J. M. & Swartz, K. J. *Neuron* **40**, 527–536 (2003).

Condensed-matter physics

Charge-ordering in oxides

Michael Coey

Transition metals form mixed-valence oxides that are expected to have ordered arrangements of *d*-shell electrons. But the ionic picture must be rethought to include oxygen 'holes' in the charge-ordered patterns.

The crystal chemistry of oxides seemed as solid as a rock; indeed, most rocks are oxides, in which 90% of the ions involved — Si^{4+} , Al^{3+} , Mg^{2+} , Na^+ , O^{2-} — adopt a stable configuration with just ten electrons. To form ionic bonds, the electrons transfer their allegiance from metals to oxygen, filling up the $2p$ shell to arrive at a $1s^2 2s^2 2p^6$ configuration. This transfer in oxides of transition metals possessing $3d$ shells can lead to multiple charge states for the metal cations. For example, TiO_2 contains Ti^{4+} ($3d^0$), but Ti_2O_3 contains Ti^{3+} ($3d^1$) and Ti_4O_7 has a mixture of both, in the formula $\text{Ti}_2^{3+}\text{Ti}_2^{4+}\text{O}_7$. At low temperature, the extra electrons associated with Ti^{3+} should be driven by their mutual Coulomb repulsion into some sort of ordered arrangement. But, in *Physical Review B*, Herrero-Martín *et al.*¹ throw doubt on physicists' favourite model of electrons in solids by calling cation charge-ordering into question, at least in manganites. The manganites are a family of compounds that have provided harmless pleasure to generations of solid-state physicists and chemists since their discovery at the Philips Research Laboratories in the Netherlands more than 50 years ago.

Charge-ordering in solids was first envisaged by Eugene Wigner in the late 1930s. The concept was applied by E. J. W. Verwey to the transition that occurs in magnetite (Fe_3O_4) at 120 K, now known as the Verwey transition. Verwey suggested that the electrons belonging to Fe^{2+} and Fe^{3+} ions order themselves over octahedral sites coordinated by oxygen². Magnetite is nature's most common mixed-valence mineral. The charge-ordered state below a Verwey transition is less conducting than the disordered state above it, where electrons can hop or tunnel from one cation site to the next. The charge order-disorder transition is driven, at least in part, by the entropy of disorder.

The mixed-valence manganites³ are solid

solutions — compositionally random mixtures of two perovskite-structure oxides that have the general formula ABO_3 . Examples include LaMnO_3 , for which the presence of the La^{3+} charge state implies a $\text{Mn}^{3+} 3d^4$ configuration; and CaMnO_3 , for which the Ca^{2+} charge state implies the smaller $\text{Mn}^{4+} 3d^3$ configuration. Many rare-earth elements can replace La, and other alkaline earths can replace Ca — providing a large chemical playground that filled up in the mid-1990s with researchers who had become bored with the structurally related superconducting cuprates. The manganites offered a bewildering profusion of crystallographic, magnetic and electronic ground states, as well as catchy physical phenomena such as magnetic melting of charge order and colossal magnetoresistance — where the resistance of the material drops dramatically as a magnetic field is applied.

The 50–50 solid solutions, such as $(\text{La}_{0.5}\text{Ca}_{0.5})\text{MnO}_3$, seemed relatively straightforward. They undergo two transitions on

cooling, the first at about 220 K to a ferromagnetic — spontaneously magnetic — conducting state, and the second near 160 K to an insulating antiferromagnetic state. At this second transition, the Mn^{3+} and Mn^{4+} ions were thought to adopt the charge-ordered checkerboard arrangement⁴ shown in Fig. 1. Another wrinkle is that the fourth $3d$ electron can occupy one of two different orbitals, a wasp-waisted cigar and a four-leaved clover. A pattern of orbital order may exist independently of charge order⁵. The octahedral sites occupied by $3d^4$ ions tend to distort in response to the orbital shape so as to lower the electrostatic energy. This site distortion is known as the Jahn–Teller effect.

Herrero-Martín *et al.*¹ have probed the local structure around the manganese ions in the 50–50 solid solution $(\text{Nd}_{0.5}\text{Sr}_{0.5})\text{MnO}_3$ by using resonant scattering of an intense synchrotron-generated X-ray beam. They do indeed find two manganese sites, and one of them is distorted as expected. However, the charge difference between the two configurations is only 0.16 electrons¹ — a far cry from $3d^3$ and $3d^4$, but in line with other experimental^{6–10} and computational^{11,12} evidence that there is little difference in manganese charge or magnetic moment on different sites. An alternative model is the Zener polaron⁹, in which pairs of Mn^{3+} cations share a 'hole' on a bridging O^- ion. These oxygen holes may form a soft charge-density wave^{10,12}.

Doubts about charge-ordering are also circulating for magnetite, the archetypal mixed-valence oxide. García and Subías argue in a topical review¹³ that no good evidence for distinct Fe^{2+} and Fe^{3+} charge states below the Verwey transition has ever been found. This is probably an overstatement; nuclear magnetic resonance and Mössbauer probes of the iron hyperfine fields reveal many distinct components below 120 K that are associated with more than two charge configurations, intermediate between

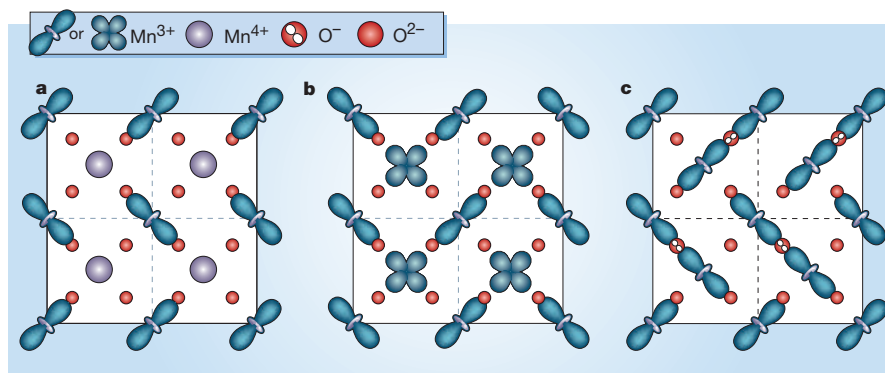


Figure 1 Called to order. a, The checkerboard charge-ordered arrangement of Mn^{3+} and Mn^{4+} ions, originally proposed for $(\text{La}_{0.5}\text{Ca}_{0.5})\text{MnO}_3$ by Goodenough⁴. b, A pattern of orbital order for Mn^{3+} ions⁷, which implies that there is incomplete occupancy of the oxygen $2p$ shell. c, The ordered arrangement of O^- ions between Mn^{3+} pairs in the Zener polaron model^{9,11}. Four of the six oxygen ions forming an octahedral cage around each of the manganese sites are shown; the other two lie above and below the plane.

Fe^{2+} and Fe^{3+} . What seems certain is that a belief in *integer 3d* configurations in a charge-ordered state of a mixed-valence oxide is unsustainable. The ionic picture is really only a starting point; highly charged states such as Mn^{4+} are mitigated by covalence, and the bonding electrons are shared in states that cannot be assigned solely to metal or oxygen. The intact $2p^6$ shell is an illusion.

The problem is to deal with on-site and intra-site electron correlations in more realistic terms. The Hubbard model — beloved of condensed-matter theorists and in which oxides are modelled in terms of cation charge fluctuations with an on-site Coulomb interaction, ignoring oxygen states — seems ever more schematic. An easy way of coming to terms with oxygen holes is needed; simple ideas have the power to guide us even when they are only half-truths. The stock response that everything is more complicated than we first thought is never much help. The entropy of the transitions is evidence that something is disordered. But what? Perhaps it will prove possible to combine the Zener polaron and cation charge-order pictures. Computational analysis of realistic crystal structures¹⁴ offers the prospect of resolving the perennial

‘chicken and egg’ question of cause and effect that arises whenever an electronic phase transition accompanies a lattice distortion. ■

Michael Coey is in the Department of Physics, Trinity College, Dublin 2, Ireland.

e-mail: jcoey@tcd.ie

1. Herrero-Martín, J., García, J., Subías, G., Blasco, J. & Concepción Sánchez, M. *Phys. Rev. B* (in the press); preprint at <http://arxiv.org/abs/cond-mat/0406407> (2004).
2. Verwey, E. J. W. *Nature* **144**, 327–328 (1939).
3. Coey, J. M. D., Viret, M. & von Molnar, S. *Adv. Phys.* **48**, 167–293 (1999).
4. Goodenough, J. B. *Phys. Rev.* **100**, 564–573 (1955).
5. Khomski, D. in *Spin Electronics* (eds Ziese, M. & Thornton, M. J.) 89–116 (Springer, Berlin, 2001).
6. Radaelli, P. G., Cox, D. E., Marezio, M. & Cheong, S.-W. *Phys. Rev. B* **55**, 3015–3023 (1997).
7. Grenier, S. *et al. Phys. Rev. B* **69**, 134419 (2004).
8. García, J., Concepción Sánchez, M., Blasco, J., Subías, G. & Grazia Proietti, M. *J. Phys. Condens. Matter* **13**, 3243–3256 (2001).
9. Daoud-Aladine, A., Rodriguez-Carvajal, J., Pinsard-Gaudart, L., Fernandez-Diaz, M. T. & Revcolevschi, A. *Phys. Rev. Lett.* **89**, 097205 (2002).
10. Loudon, J. C. *et al.* preprint at <http://arxiv.org/abs/cond-mat/0308581> (2003).
11. Zheng, G. & Patterson, C. H. *Phys. Rev. B* **67**, 220404 (2003).
12. Ferrari, V., Towler, M. D. & Littlewood, P. B. *Phys. Rev. Lett.* **91**, 227202 (2003).
13. García, J. & Subías, G. *J. Phys. Condens. Matter* **16**, R145–R178 (2004).
14. Patterson, C. H. Preprint at <http://arxiv.org/abs/cond-mat/0405299> (2004).

Immunology

Another manifestation of GOD

Martin F. Flajnik

In studies of the evolution of the adaptive immune system, the lamprey has been an unlikely centre of attention. These studies now provide evidence of a fascinating variation on how such a system can operate.

Like almost all other vertebrates, we humans have jaws. Deep in vertebrate family history, however, are our jawless relatives, most of which are extinct. The exception is the lamprey, which, along with its cousin the hagfish, is alive and well, and has posed a real puzzle for immunologists: lampreys seem to get by without one of the two arms of jawed-vertebrate immune defence, the so-called adaptive system, which is well developed in the oldest jawed vertebrates, the sharks. As Pancer *et al.* show on page 174 of this issue¹, however, lampreys apparently have such a system. It's just rather different.

The other arm of the immune system is the innate system, which reacts swiftly to invaders with a generalized response mounted by receptors encoded in the germ line. The more glamorous adaptive system is defined by non-germline-generated receptors that bind an invader's signature molecules, or antigens, and can recognize any biological molecule in nature. These antigen receptors — antibodies or T-cell receptors — are generated by a gene-rearrangement process during lymphocyte development². As shown in

Fig. 1a, the antigen-binding or variable (V) region is formed by the rearrangement of V, D (diversity) and J (joining) gene segments for one chain of the receptor, and of V and J segments for the other chain (for antibodies' heavy and light chains, respectively). The enzymes RAG1 and RAG2 stitch these segments together³. The multigenic nature and imprecise joining of the V, D and J segments result in each individual having a huge repertoire of antigen receptors.

Immunologists irreverently refer to this process as GOD (generation of diversity). For such a system to be efficient, the antigen receptors must be expressed in clones of lymphocytes⁴ — B cells for antibodies and T cells for T-cell receptors. Only those lymphocytes that can bind to a particular antigen are stimulated to secrete anti-pathogen antibodies (B cells) or produce 'effector' T cells that can, directly or indirectly, kill our own cells infected by intracellular invaders.

Evidence for the existence of this adaptive system in invertebrates and in lampreys and hagfish has been pursued valiantly, but in vain. The failed search for antibodies and

T-cell receptors — as well as for other components of the adaptive system, such as genes encoding the major histocompatibility complex (MHC) important for T-cell stimulation, the tissues where the lymphocytes are generated and then stimulated by antigen, and the RAG genes themselves — has led to the premature end of several graduate student careers. We in the comparative immunology field now think it unlikely that such apparatus will be uncovered outside the jawed vertebrates⁵. The unusual nature of the RAG genes, and the recombination signal sequences that they recognize (Fig. 1a, overleaf), have led to a 'big bang' hypothesis⁶, in which mobile genetic elements, including the RAG genes, fortuitously invaded and disrupted an ancestral antibody–T-cell-encoding genomic sequence⁷. These sequences could then be stitched together only by RAG action, resulting in a mechanism for GOD and the conception of the jawed-vertebrate adaptive immune system.

Now wait one minute — all this does not mean that GOD could not be realized in another way in jawless vertebrates or invertebrates. Indeed, the same end can be achieved by mutational or gene-conversion processes in some jawed vertebrates, and such mechanisms could have preceded the RAG invasion⁸. Furthermore, GOD in jawed vertebrates is not the only prototype for gene rearrangement in the animal kingdom: parasites have long been known to have mobile genes that create rapidly evolving proteins capable of tricking the host's immune system⁹.

Rather than assuming that if lampreys had an adaptive immune system it would (perhaps) be a forerunner of the vertebrate system, Pancer *et al.*¹ wondered whether any adaptive response would be carried out by lymphocytes. So they performed a search for lymphocyte-specific genes. In earlier work in collaboration with another laboratory, this group purified lamprey cells on the basis of their lymphocyte-like properties and searched for expressed sequence tags (ESTs) within these cells¹⁰. Like the rest of us, they found no antibodies, T-cell receptors, MHC and so on. But they did unearth ESTs for transcription factors, cell-surface molecules and other proteins that are expressed more or less specifically in jawed-vertebrate lymphocytes. They also found unique ESTs for leucine-rich repeats (LRRs), which are typical of many proteins involved in immune defence¹¹.

Pancer *et al.*¹ have now shown that these genes are upregulated in antigen-stimulated lymphocytes. Additionally, they found that this gene family is extremely diverse in the number and types of LRR 'cassettes' encoded by each complementary DNA clone, and therefore dubbed them 'variable lymphocyte receptors' or VLRs. This diversity could have been the result of either a vast number of

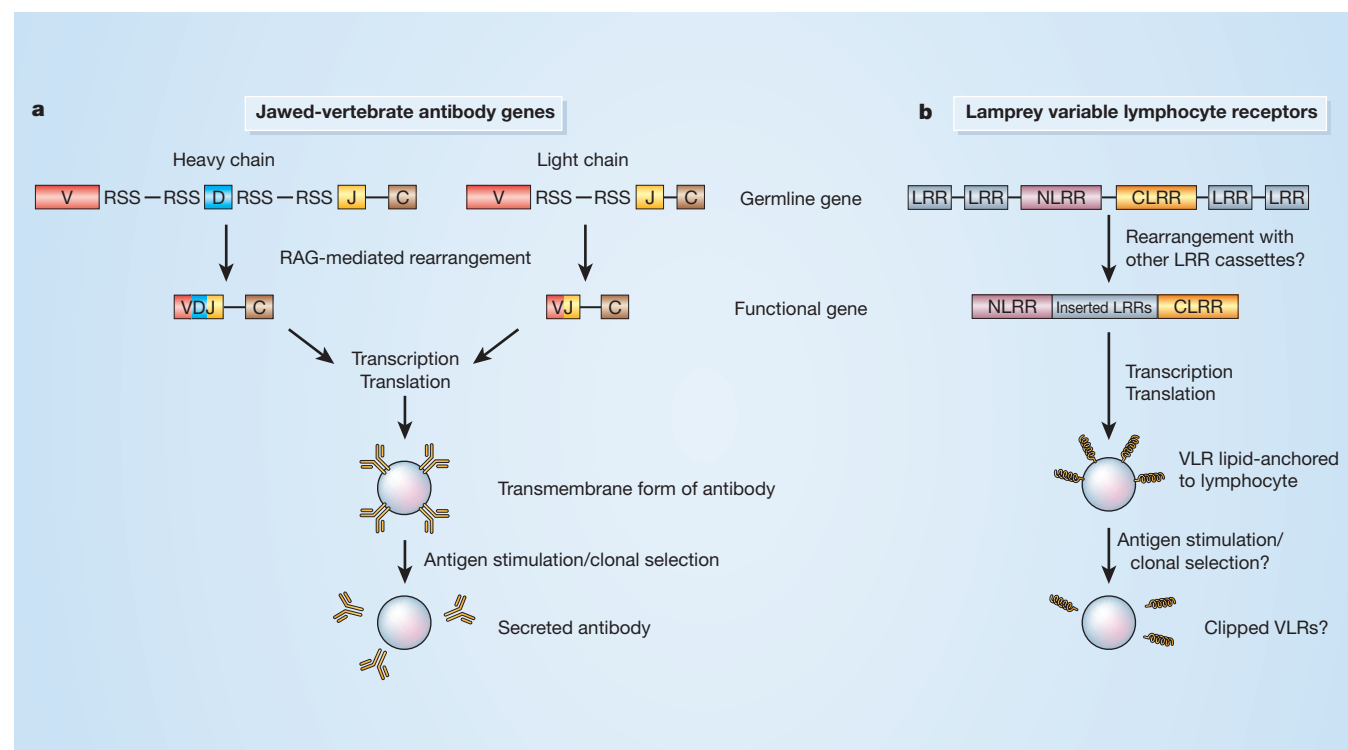


Figure 1 Adaptive immunity in jawed vertebrates and the jawless lamprey. a, Antibody production as the example for jawed vertebrates. During lymphocyte development, the RAG enzymes rearrange genes by recognizing recombination signal sequences (RSS) that flank the variable (V), diversity (D) and joining (J) gene segments at heavy-chain and light-chain loci (C, constant regions). These segments are stitched together by enzymes, and the functional gene encodes the binding sites that can recognize foreign antigen. A huge variety of such sites can be produced. But only one type of antibody is expressed by each lymphocyte; thus, only cells expressing antibodies specific for a particular antigen will

become stimulated, proliferate by clonal selection and secrete antibodies to destroy an invader. b, Variable lymphocyte receptors (VLRs) in the lamprey¹. The VLR genetic locus consists of segments at each end of the functional gene that encode leucine-rich repeats (NLRR or CLRR). By an unknown mechanism, other LRR gene cassettes from upstream and downstream of these segments are inserted to produce a functional gene. Preliminary indications are that each lamprey lymphocyte has a specific VLR, implying that clonal selection also operates in this system. VLRs are probably anchored to the lymphocyte membrane and released after antigenic stimulation.

germline genes, as for other innate receptors¹², or of non-germline rearrangement. Pancer *et al.* found that invariant LRR cassettes were encoded by each end of each clone of complementary DNA, but that LRRs sandwiched between them varied in number (Fig. 1b). Inspection of genomic DNA showed the invariant cassettes to be encoded by single-copy genes near to each other in non-lymphocyte DNA. In further analyses of lymphocyte genomic DNA, it emerged that the other LRR cassettes were somehow added and sewn together. Furthermore, preliminary evidence suggests that each lamprey lymphocyte bears a unique VLR, setting the stage for clonal selection reminiscent of the RAG-based system.

This provocative discovery raises many stirring questions. How is GOD accomplished for the VLR? The LRR gene cassettes are not flanked by signal sequences recognized by RAG proteins, so the VLR diversity-generating mechanism has probably arrived at the same general principle as the antibody–T-cell-receptor system from different origins. Can the VLR account for old experiments that suggested adaptive immunity in jawless fish³? This is possible, but because the VLR are lipid-anchored to the lymphocytes

they would have to be clipped off to be secreted into body fluids and bind to foreign invaders, a totally different mechanism from that for the initially transmembrane antibodies in jawed vertebrates.

Is this an old system that was superseded by jawed-vertebrate immunity, or is it something that arose only in a lamprey ancestor? The evolutionary split between lampreys and hagfish occurred more than 300 million years ago; if hagfish (and perhaps some invertebrates or even jawed vertebrates as well) have VLRs, it suggests that the ‘rules’ for the generation of an adaptive immune system arose early in evolution, and indeed that the VLR system was superseded by our current adaptive system. If VLRs are the foundation of the lamprey adaptive immune system, do the lymphocytes expressing them form a single subset? Or, as in the jawed vertebrates, are there subsets of cells for different functions?

Revelations of other immune diversity-generating mechanisms may be on the way. Data from invertebrates such as echinoderms and molluscs (L. C. Smith, George Washington Univ.; E. Loker, Univ. New Mexico, personal communications) suggest that non-germline mechanisms for GOD might

abound in the animal kingdom, especially among animals that are relatively long-lived or complex, or both. The jawed-vertebrate adaptive immune system, based on RAG enzymes, was a magnificent and unique innovation. But Pancer *et al.* have shown that the RAG-based system is not the be-all and end-all of adaptive immunity.

Martin F. Flajnik is in the Department of Microbiology and Immunology, Room 13-009, University of Maryland, 655 West Baltimore Street, Baltimore, Maryland 21201, USA.
e-mail: mflajnik@som.umaryland.edu

1. Pancer, Z. *et al.* *Nature* **430**, 174–180 (2004).
2. Hozumi, N. & Tonegawa, S. *Proc. Natl Acad. Sci. USA* **73**, 3628–3632 (1976).
3. Schatz, D. G., Oettinger, M. A. & Baltimore, D. *Cell* **59**, 1035–1048 (1989).
4. Jerne, N. K. *Proc. Natl Acad. Sci. USA* **41**, 849–857 (1955).
5. Flajnik, M. F., Miller, K. & Du Pasquier, L. in *Fundamental Immunology* 5th edn (ed. Paul, W. E.) 519–570 (Lippincott Williams & Wilkins, Philadelphia, 2003).
6. Bernstein, R. M., Schluter, S. F., Bernstein, H. & Marchalonis, J. J. *Proc. Natl Acad. Sci. USA* **93**, 9454–9459 (1996).
7. Thompson, C. B. *Immunity* **3**, 531–539 (1995).
8. Lee, S. S. *et al.* *Immunity* **16**, 571–582 (2002).
9. Donelson, J. E. *Acta Trop.* **85**, 391–404 (2003).
10. Uinuk-Ool, T. *et al.* *Proc. Natl Acad. Sci. USA* **99**, 14356–14361 (2002).
11. Bell, J. K. *et al.* *Trends Immunol.* **24**, 528–533 (2003).
12. Hawke, N. A. *et al.* *Proc. Natl Acad. Sci. USA* **98**, 13832–13837 (2001).

Fluid mechanics

Different strokes for nano-folks

Phys. Rev. E **69**, 062901 (2004)

Swimming isn't easy when you're small. Humans can propel themselves through water using a cyclic set of motions, thanks to inertia following the forward stroke. But at the microscopic scale, inertia is negligible and viscosity dominates. In this situation, a cyclic sequence that looks the same backwards and forwards in time gets you nowhere — the second half of the cycle cancels the first. That's why a scallop can't move by simply opening and closing its shell. So micro-swimmers have to be more inventive with their strokes.

Ali Najafi and Ramin Golestanian have devised a sequence of moves that should propel a very simple microscopic swimming device. The 'swimmer' is a mechanism of three spheres connected in a linear arrangement by two piston-like arms. The middle sphere has a power source that extends or contracts the two arms. First, the rear arm contracts; then the forward arm does the same; then the rear arm extends, followed by the front arm. A simple analysis shows that this time-asymmetric sequence should push and pull the device through a viscous fluid.

But will it really work? Making such a tiny device, using either microfabrication technology or, at an even smaller scale, molecular assemblies that can produce mechanical displacement, is a daunting but not unrealistic challenge, the authors believe.

Philip Ball

Chemistry

On the other hand...

Angew. Chem. Int. Edn **43**, 3317–3321 (2004)

Why are amino acids left-handed? Non-biological chemistry generally produces equal numbers of left- and right-handed molecules, called enantiomers, yet life chooses to use just one form to build proteins. One explanation is that, as the first chemical reactions of life began, a minute initial imbalance in enantiomers underwent some reaction that left them predominantly left-handed. The only example to date of such a process, reported by Kenso Soai in 1995, relies on chemistry that would be impossible in the conditions that are thought to have prevailed on the prebiotic Earth.

Suju P. Mathew *et al.* now report analyses of another reaction that can pull itself up by its bootstraps in this way, one that is more compatible with the conditions on the early Earth. It not only generates an organic product that speeds up its own reaction, but

also increases the proportion of left-handed versions of itself in the mixture.

Tantalizingly, this reaction relies on an amino-acid catalyst, proline, as a source of asymmetry. The authors believe that proline and the reaction product bind together to make a more effective catalyst. They hope that a better understanding of the reaction mechanism will in time enable them to produce a truly autocatalytic system, in which the reaction product itself acts as the catalyst.

Mark Peplow

Natural hazards

Houses on fire

Geophys. Res. Lett. **31**, L12212 (2004)

In regions that are prone to wildfires, the edge of town can clearly be a risky place for human habitation. But just how risky? Putting numbers to fire hazard is essential for developing rational planning regulations (how far from the bushland should houses be built?) and fair and realistic insurance premiums.

Keping Chen and John McAneney have used aerial and satellite images of bushfire damage in Duffy, Canberra (2003), and Como-Jannali, Sydney (1994), to deduce how far such fires penetrate into urban settlements. They find that homes more than 700 metres from the adjacent bushland are generally safe, and that the percentage of houses burned down decreases linearly with distance from the edge of the bush.

In both these and other cases of severe bushfires, around 60% of the homes within the first 50 metres were destroyed — a remarkably consistent figure, given the varying circumstances of the fires. Chen and McAneney conclude that most of the houses were set ablaze by wind-borne embers rather than by radiant heat or direct contact with the fire.

Philip Ball



Duffy, Canberra, on fire in 2003: bushfires threaten the outskirts of towns and cities.

Human evolution

Small faces

J. Hum. Evol. **46**, 655–677 (2004)

There is evidence that prehistoric humans had larger facial features than we have today. But why should this be? One theory is that the advent of cooking and other forms of food processing lessened the workload on our chewing apparatus, in turn reducing the amount of facial growth that is stimulated during development. Research showing the influence of diet on growth in a small mammal now lends support to this idea.

Daniel E. Lieberman and colleagues studied the development of the rock hyrax (*Procapra capensis*), bizarrely the closest living relative of elephants. Like humans, the rock hyrax's molars are directly beneath its eye sockets, so that its chewing style should resemble our own in some respects.

From the age of five to six months, the animals were either given a diet of raw and dried food, or fed the same diet cooked until soft. After a further three months, those raised on soft food showed around 10% less growth in certain facial areas. The authors believe that the difference between the two groups mirrors the differences we see between modern humans and our ancestors raised on less processed diets.

Michael Hopkin

Developmental biology

Primitive streak singled out

Development **131**, doi:10.1242/dev.01178 (2004)

During the early stages of a chick embryo's growth, any part of the embryo has the ability to form a 'primitive streak', which defines which part will ultimately form the head and which the tail. Federica Bertocchini *et al.* now show why, during normal development, only one such structure typically forms.

On one side of a chick embryo, the researchers implanted a ball of cells that make Vg1, a protein that triggers formation of the primitive streak. Some time later, they implanted a second identical ball on the opposite side.

When the second ball of cells was inserted 6 hours or more after the first, it was unable to trigger a second primitive streak. This suggests that Vg1 switches on production of an inhibitor protein, which speeds across the 3-mm embryo at 0.5 mm per hour — up to ten times as fast as comparable molecules — and shuts down primitive-streak formation elsewhere.

Bertocchini *et al.* have yet to identify the inhibitor protein, but they believe that a similar mechanism lays down the body axis of mammalian embryos. They speculate that identical or conjoined twins may sometimes form because of problems with the inhibitor or the system that transports it across the embryo.

Helene Pearson

The RNAi revolution

Carl D. Novina and Phillip A. Sharp

The term RNAi — short for RNA interference — crops up again and again in biology research these days. This is in part because of its power as a laboratory tool, and in part because it is a widespread natural phenomenon.

Arguably the most important advance in biology in decades has been the discovery that RNA molecules can regulate the expression of genes. For years, RNAs were thought to have just two broad functions in cells. Single-stranded messenger RNAs (mRNAs) are vital intermediaries in gene expression, transmitting information between DNA and protein. Ribosomal and transfer RNAs have structural, catalytic and information-decoding roles in the process of protein synthesis.

This picture became complicated in 1998, however, when Andrew Fire, Craig Mello and colleagues¹ announced their discovery of RNA interference (RNAi) — the silencing of gene expression by double-stranded RNA molecules — in nematode worms. Since then it has become clear that RNAi is an extremely useful experimental tool for learning what genes do. Not only that, it is an evolutionarily ancient method of genome defence in many organisms. Research on the mechanistic side has also proceeded at a staggering pace. Double-stranded RNA was originally found to silence genes by marking out their mRNA intermediaries for destruction. But at least two more mechanisms have been uncovered: blocking transcription (the synthesis of mRNAs from DNA) and inhibiting translation (the synthesis of proteins from mRNAs). And this is surely not the end of the story.

Discovery

Like many significant advances, the discovery of RNAi was technically simple: injecting double-stranded RNAs (dsRNAs) into the worm *Caenorhabditis elegans* resulted in the silencing of a gene whose sequence was complementary to that of the dsRNAs¹. All discoveries have precedents, however, and RNAi was no exception. In the late 1980s and early 1990s, plant biologists working with petunias were surprised to find that introducing numerous copies of a gene that codes for deep purple flowers led, not as expected, to an even darker purple hue, but rather to plants with white or patchy flowers^{2,3}. Somehow, the introduced genes (the 'transgenes') had silenced both themselves and the plants' own 'purple-flower' genes. Similarly, when plants were infected with an RNA virus that had been genetically engineered to contain fragments of a plant gene, the plant's gene itself became silenced⁴.

At the time these observations were

Short interfering RNAs (siRNAs)	A class of double-stranded RNAs of 21–22 nucleotides in length, generated from dsRNAs. siRNAs silence genes by promoting the cleavage of mRNAs with exactly complementary sequences, or recruiting inhibitory proteins to, or directing the modification of, DNAs with exactly complementary sequences.
MicroRNAs (miRNAs)	A class of 19–25-nucleotide, single-stranded RNAs that are encoded in the genomes of most multicellular organisms studied. Some are evolutionarily conserved and are developmentally regulated. They silence certain cellular genes at the stage of protein synthesis.
Tiny non-coding RNAs (tncRNAs)	A newly discovered class of short, 20–22-nucleotide RNAs that are encoded in the genome of <i>C. elegans</i> . They are not evolutionarily conserved, but some are developmentally regulated. Their function is still unknown.
Small modulatory RNA (smRNA)	A short, dsRNA, identified earlier this year in mice, that allows the expression of neuron-specific genes only in adult neurons.

Figure 1 The short story of gene silencing — the main characteristics of short RNAs.

puzzling, but they were readily understood when Fire and Mello's work was published. RNA viruses replicate through dsRNA intermediates; multi-copy transgenes can produce low levels of dsRNAs as well. The results hinted that RNAi, mediated through dsRNAs, is an ancient process, which predates the evolutionary divergence of plants and worms. RNAi thus naturally functions, it is thought, to silence viruses and rogue genetic elements that make dsRNA intermediates — types of RNA not usually produced by cells.

Mechanisms

So how does this silencing process work? An early step in determining the mechanism was the observation that plants that were silencing genes in an RNAi-related process (dubbed post-transcriptional gene silencing) produced short RNAs, some 20–25 nucleotides long, that matched the gene being silenced⁵. These RNAs are much shorter than typical mRNAs and ribosomal RNAs, perhaps explaining why this fundamental process remained undiscovered until recently.

At around the same time, the biochemical reactions of RNAi were reconstituted *in vitro* using fruitfly extracts^{6,7}. These studies revealed that the long dsRNAs were diced up into short RNAs with a specific structure: two 21-nucleotide strands of RNA in a staggered duplex, with 19 nucleotides of dsRNA and two unpaired nucleotides at the ends⁸. This species was dubbed a short interfering RNA (siRNA; Fig. 1), and synthetic versions were shown to silence genes⁹.

The enzyme responsible for cleaving

dsRNAs into siRNAs is called Dicer¹⁰. In the next step in the process, the 'sense' strand of an siRNA — the strand that has exactly the same sequence as a target gene — is removed, leaving the 'antisense' strand, which is complementary to the target gene, to function in gene silencing (Fig. 2). The antisense strand guides a protein complex to the mRNA produced from the target gene, destroying it.

Early studies of transgene silencing in plants indicated that an RNAi-type process also works at the level of transcription (in a process called transcriptional gene silencing; reviewed in ref. 11). Further studies showed that both strands of a transgene are probably transcribed (rather than just one strand, as is the case for typical genes), generating a dsRNA rather than a single-stranded mRNA. The dsRNA is then processed into siRNA, which somehow initiates transcriptional silencing (Box 1). Simply put, this seems to involve the reversible addition of chemical groups to nuclear DNA and the proteins that package it as dictated by the sequence of the dsRNA¹¹.

As might be expected, there are nuances in the process of RNAi. For instance, in some organisms, gene silencing by RNAi requires the activity of an RNA-dependent RNA polymerase (RdRP), which uses the antisense strand of an siRNA as a primer with which to make more dsRNAs, thereby amplifying the process (Fig. 2). In plants, this amplification enables RNAi-mediated gene silencing to spread throughout non-reproductive (somatic) tissues, by cell-to-cell transfer of dsRNAs — generating widespread resistance

to viral infections. Amplification is not, however, thought to occur in vertebrates or fruitflies, as genes for a RdRP have not been found in these organisms.

Functions

Short interfering RNAs may have a general role in the silencing of transposable elements ('jumping genes'), repetitive genes (including transgenes) and viruses. About half of the sequences in the human genome, for instance, were generated by duplication and insertion of transposons — parasitic elements that create 'junk' DNA. Silencing of these mobile elements is crucial for the stability of the genome, and RNAi-related processes involving the generation of short RNAs are essential to this silencing process in many organisms.

For example, worms with mutations in genes of the RNAi pathway are unable to silence transposons in germline tissues (those that produce eggs and sperm)^{12,13}. Similarly, plants with mutations in subsets of RNAi-pathway genes show defects in the silencing of specific mobile genetic elements¹⁴.

The same process could also limit the degree to which a gene can be expressed in certain tissues. For example, fruitflies show 'dosage-dependent' gene silencing in a process related to RNAi¹⁵. Here, more gene copies result in higher gene expression — but only up to a point, beyond which gene expression is reduced. In this and similar cases, the proteins that package DNA within cells are modified to suppress transcription. Similarly, in fission yeast, siRNAs directly silence the expression of repetitive transposable elements at the level of transcription¹⁶. Intriguingly, this silencing can spread to nearby regions, repressing gene regions that are adjacent to the transposable elements (Box 1).

MicroRNAs

The existence of siRNAs produced in response to experimentally introduced long dsRNAs and transgenes prompted a search for short RNAs that are actually encoded by genes in the host genome. Researchers embarked on the process of purifying short RNAs (of 19–25 nucleotides) from several animal species, and were surprised to find, not siRNAs, but a new class of short RNAs, dubbed microRNAs (miRNAs; Fig. 1).

MicroRNAs are abundant, single-stranded RNAs, varying from a few thousand to 40,000 molecules per cell¹⁷. They are encoded in the genome, but, unlike mRNAs, do not represent a step towards the production of a protein. Instead, they regulate the expression of mRNAs. They are thought to be formed from 'hairpin' RNAs — RNAs that are effectively double-stranded by virtue of the fact that they are self-complementary and fold over in the middle. Like siRNAs, miRNAs are produced by the cleavage of dsRNAs by Dicer and related activities. They have been found in all multicellular organisms studied,

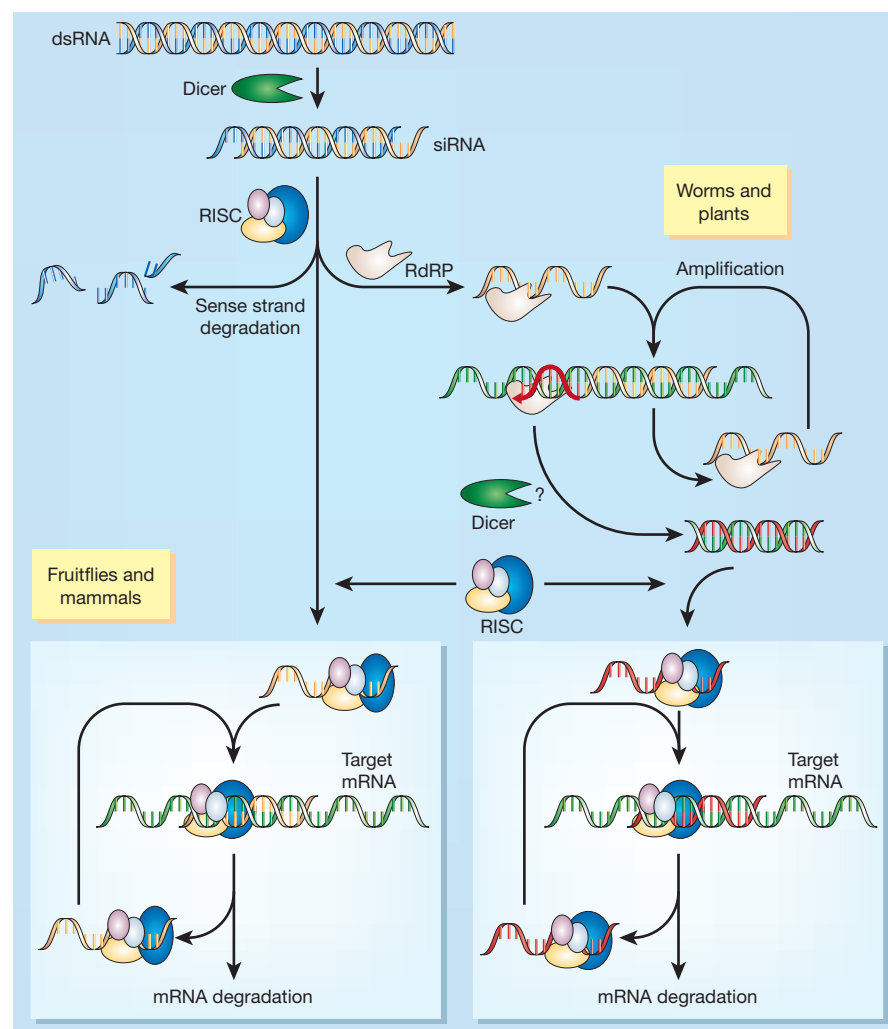


Figure 2 Short interfering RNAs — post-transcriptional gene silencing. RNAi is triggered when a cell encounters a long double-stranded RNA (dsRNA), which might be produced from an introduced transgene, a viral intruder or a rogue genetic element. An enzyme called Dicer cleaves the long dsRNA into siRNAs. An RNA-induced silencing complex (RISC) then distinguishes between the different strands of the siRNA. The sense strand (blue) is degraded. The antisense strand (yellow) is used to target genes for silencing, and has one of several fates depending upon the organism. In fruitflies and mammals, the antisense strand is incorporated directly into RISC to target a complementary mRNA (green) for destruction. In the absence of siRNAs, the RISC lacks sequence-specific mRNA-binding properties. But when bound to the antisense strand, the now activated RISC can participate in repeated cycles of degradation of specific mRNAs, such that no protein is made — effectively silencing the gene from which the mRNAs are produced. In worms and plants, the antisense strand of the siRNA might first be used in an amplification process. The antisense strand, bound by an RNA-dependent RNA polymerase (RdRP) enzyme, can pair up with a complementary mRNA (green) and act as a start point for the synthesis of a new long dsRNA. Dicer is then required to generate new siRNAs (red), which are specific to different sequences on the same mRNA. Again, the target mRNA is destroyed.

and their encoding genes seem to make up 0.5–1% of the predicted genes in these organisms — 200–255 in humans, for instance¹⁸. So far, some 175 of these human miRNA genes have been confirmed by biochemical analysis¹⁹.

On the basis of a model established from studies of nematode development, miRNAs in mature animals are thought to regulate the expression of distinct genes by base-pairing with partially complementary sequences in mRNAs and simply inhibiting their translation into proteins (Fig. 3). They therefore work by a somewhat different mechanism

from siRNAs, in that transcription is not impeded and mRNAs are not destroyed. Another crucial difference is that siRNAs generally target the genes or genetic elements from which they originated, whereas miRNAs regulate separate genes — perhaps hundreds or more per miRNA. Furthermore, the degree of translational inhibition by miRNAs is thought to depend on how many of these molecules are bound to the target mRNA. Typically, such an mRNA contains many binding sites at one end (the 3'-untranslated region), and several different miRNAs can target the same 3' region.

So what sort of biological processes do miRNAs regulate? In worms, the two founding members of the miRNA class of genes — *lin-4* and *let-7* — determine transitions between larval stages in post-embryonic development^{20,21}. In plants, miRNAs also control crucial developmental transitions, and in flies they control cell division and cell death (reviewed in ref. 22). Human miRNAs were identified only a short time ago, and little is known about their function. As miRNAs are evolutionarily conserved, however, there can be little doubt about their importance in human physiology.

It can be difficult to identify the precise mRNA targets of miRNAs; computational approaches to the problem are complicated by the imperfect base-pairing between the two. Several studies have used rules inferred from known examples of miRNA–mRNA pairing, combined with a requirement for evolutionary conservation, to predict possible targets. One such study predicted an average of around five target mRNAs per miRNA²³. Given that there are some 200 human miRNAs¹⁸, it seems likely that several thousand human mRNAs could be targeted.

Common ground

Not all miRNAs work by binding with imperfect complementarity to mRNAs and preventing their translation. When miRNAs were first cloned from plants, it was recognized that many show perfect or near-perfect complementarity to cellular mRNAs, implying that these mRNAs might instead be targeted for degradation in an siRNA mode of action. This was indeed shown to be the case. Many of the plant mRNAs that are regulated by miRNA-directed cleavage encode gene-transcription factors that are important in plant development (reviewed in ref. 22). Recently, a conserved miRNA has been found in mammals which directs the cleavage of a target mRNA and, as in plants, regulates the expression of a transcription factor that is important for normal development²⁴. Therefore, miRNA activities probably indirectly control the expression of vast numbers of genes.

Could siRNAs also work like miRNAs — inhibiting translation rather than targeting mRNAs for destruction? Our research with mammalian cells, using chemically synthesized siRNAs that were partially complementary to numerous target sites in a 'reporter' gene, showed that the siRNAs indeed inhibited the expression of the reporter at the stage of protein synthesis²⁵. The major distinction, as noted above, is that siRNAs usually silence a target mRNA efficiently (through degradation) if the mRNA contains a single site that is almost exactly complementary to the short RNA. In contrast, the miRNA mode of action requires that the mRNA contains numerous partially complementary binding sites, which function synergistically. These observations may

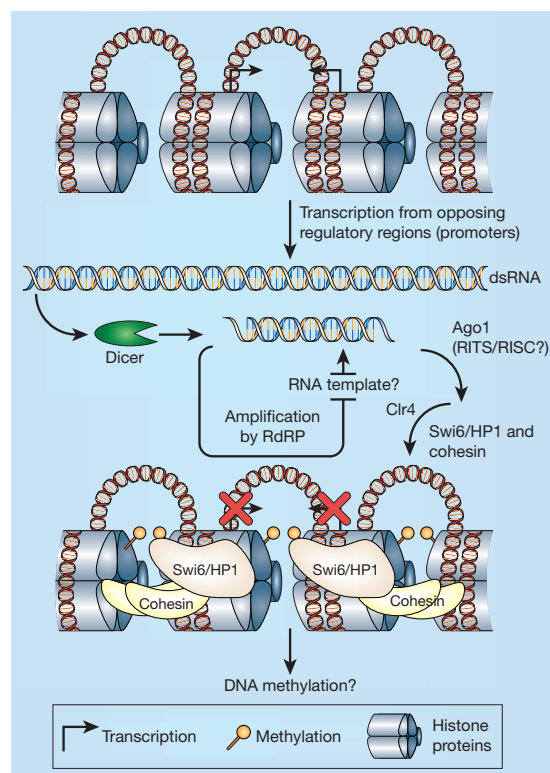
Box 1 Transcriptional gene silencing

One mechanism of action of siRNAs involves silencing gene transcription. This occurs through its effects on chromatin, the compact form of DNA, in which DNA is wrapped around complexes of histone proteins.

Long dsRNAs of similar sequence to genomic DNA might be produced in flies and mammals by symmetrical transcription from opposing promoters; this often occurs when many copies of transgenes or other repeated sequences are present, and in plants and worms through the activity of an RNA-dependent RNA polymerase (RdRP; see Fig. 1).

In plants, worms and flies, siRNAs produced from these dsRNAs can direct the formation of silenced DNA regions. As described in Fig. 1, the sense strand of the siRNA is degraded; the antisense strand is used to recruit proteins that inhibit transcription. The antisense strand identifies the correct position along the DNA sequence, by base-pairing either with DNA or with newly forming mRNA transcripts.

The details are best understood in fission yeast. Here, the Ago1 protein acts in the RNA-induced silencing complex (RISC), or in an



RNA-induced initiation of transcriptional gene silencing (RITS) complex³⁰, to distinguish the antisense strand from the sense strand. This strand helps to recruit an enzyme that adds methyl groups to histone protein H3, forming heterochromatin, the silenced form of chromatin.

The methylated histones promote the recruitment of proteins that can spread the silencing state¹⁶. In the fully repressed state of chromatin, the DNA may also be methylated by processes requiring siRNA, although this has only been seen in plants.

explain why siRNAs do not often inadvertently and potentially silence mRNAs. Statistically, any 21-nucleotide RNA could partially base-pair with many mRNAs through continuous sequences of six to eight base pairs. But if translational silencing requires that many short RNAs bind the mRNA, the miRNA mode could be as specific as the siRNA mode.

Broadening the scope

Over the past couple of years, yet more mechanisms of action of short RNAs have been discovered — and potential new classes of such RNAs have been revealed. For instance, in the ciliate *Tetrahymena thermophila*, certain siRNAs are thought to specify sites in the organism's DNA that allow it to fragment²⁶. The exact mechanism is unknown, but the discovery supports the contention that siRNAs can specifically recognize DNA sequences. DNA fragmentation in this process requires two genes whose relatives have been implicated in RNAi and related processes in other species. Ciliates lacking

either gene do not accumulate siRNAs, and do not undergo DNA fragmentation.

Another class of short, 20–22-nucleotide RNAs in *C. elegans* was described last year²⁷ (Fig. 1). These tiny non-coding RNAs (tncRNAs) do not have highly complementary sequences nearby in the genome, and so may not be processed from hairpins, although most tncRNAs seems to be produced by Dicer. Unlike many miRNAs, the sequences of tncRNAs are not commonly conserved among related species. Although their functions are unknown, many are complementary to mRNAs and might target them for degradation.

Finally, earlier this year, a short, non-protein-coding, 20-nucleotide dsRNA, called a small modulatory RNA (smRNA), was isolated from neural stem cells in the hippocampus of adult mice²⁸. This dsRNA has the same sequence as a 21–23-nucleotide stretch of DNA called the neuron-restrictive silencer element. This is evolutionarily conserved in vertebrates and is bound by a regulatory protein

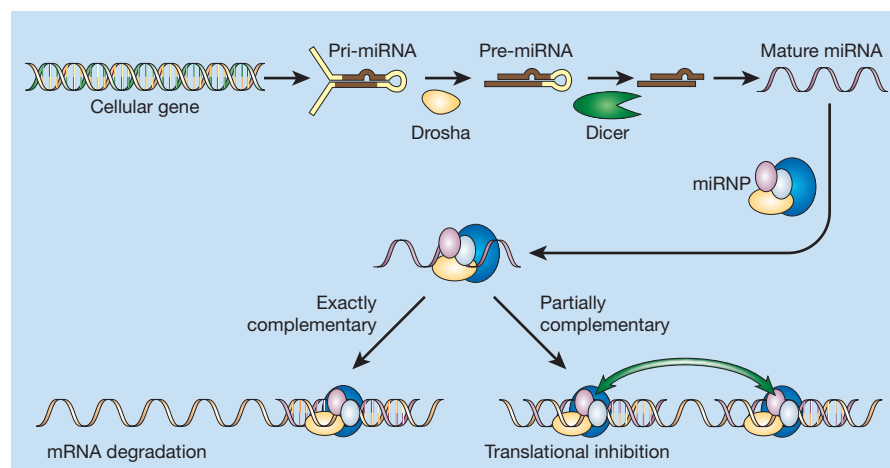


Figure 3 MicroRNAs — translational silencing. miRNAs are encoded in the genomes of all multicellular organisms studied so far. They can be expressed either from an intergenic cluster, or from single genetic regions. Both types fold into long hairpin RNAs called primary microRNAs (pri-miRNAs), which show imperfect internal sequence complementarity (the bulge represents a region where there is no complementarity). A nuclear enzyme, Drosha, cleaves pri-miRNAs into smaller, roughly 70-nucleotide hairpin RNAs termed precursor miRNAs (pre-miRNAs)³¹. These are exported from the nucleus to the cytoplasm, where another enzyme, Dicer, cleaves them into mature, single-stranded miRNAs, 21–22 nucleotides long. A mature miRNA is then assembled into a ribonucleoprotein (miRNP) complex. This complex most commonly binds to the 3'-untranslated sequences of particular mRNAs through partially complementary sequences, and prevents the mRNAs from being translated into protein. Base-pairing with the 7–8 nucleotides near the 5' terminus of the miRNA is essential here. Numerous miRNP complexes bind to the 3'-untranslated sequences of mRNAs and cooperate for effective translational silencing through interactions (green arrow) that are still unclear. In a few cases, the miRNA is exactly or nearly exactly complementary to a site in an mRNA, and this results in cleavage and degradation similar to that observed with siRNAs (see Fig. 1). This has led to the proposal that siRNAs may function by being recognized by the miRNP complex. There is genetic evidence in plants and worms, however, that different members of an extended family of proteins are required for the miRNA and siRNA pathways — so in these organisms the two pathways may be distinct.

that helps to prevent neuron-specific genes from being expressed in non-neuronal cells. In neuronal cells, where the neuronal genes are expressed, the smRNA interacts with the regulatory protein, and promotes gene expression and neuronal development. It is unclear exactly how smRNA acts, but it will be interesting to see whether it is the founding member of a new class of short RNAs.

Into the lab and clinic

RNAi promises to be one of the most useful laboratory tools yet. Classical genetic analysis of gene function involves identifying an organism with an abnormal set of physical and behavioural characteristics, and then isolating the underlying mutant genes. Typically, this 'forward genetic' approach is limited to rapidly reproducing organisms. Genetic experiments in mammals have been limited to 'reverse genetic' approaches, which begin with knocking out a specific gene to identify its function. But many of these approaches are laborious and expensive. In contrast, silencing genes with RNAi makes it possible to analyse many gene networks in mammals, and offers a higher-throughput approach to the analysis of model organisms such as *C. elegans*.

In *C. elegans*, the experiments are very simple: inject worms with dsRNA, soak them in a solution of dsRNA or feed them bacteria

expressing dsRNA, and the target gene is silenced! There are some 19,000 genes in worms, and libraries of 12,000 different dsRNAs have been used to screen this animal for genes involved in complex phenomena such as obesity and ageing. Similarly, purified long dsRNAs have been used to define the roles of specific genes in cholesterol metabolism and heart formation in fruitflies.

The use of chemically synthesized siRNAs has also helped to define gene function in vertebrate cells⁹, and has even been extended to human cells *in vitro*. Human cells are usually killed by long dsRNAs. Short RNA segments (of 21–22 base pairs) do not activate this killing process, but still possess enough sequence complexity to specify the silencing of a particular gene⁹. Libraries of short RNAs, or of DNA vectors encoding short RNAs, have been generated to target some 8,000 of the roughly 35,000 human genes to help determine their functions. Such screens have already identified gene activities in diseases such as cancer. Short RNAs might even be used to treat such diseases. A major obstacle to therapeutic gene silencing is the 'delivery problem' — the necessity of introducing short dsRNAs into specific organs. Emerging evidence suggests, however, that this obstacle might be surmountable.

Speculations

What biological phenomena might arise from RNAi-based gene regulation? In theory, any 21–22-nucleotide sequence in the genome could be converted into a potential regulatory factor by transcription into dsRNA. Consider the following series of events: first, a mutation generates an inverted duplication of a genomic sequence; second, transcription produces a hairpin RNA; and third, Dicer produces an siRNA or miRNA. These events would create a factor that could regulate another gene at the level of mRNA degradation, translation or, possibly, transcription. This series of events seems more likely than that required to create new regulatory possibilities with sequence-specific RNA- or DNA-binding proteins. Thus, we might expect more evolutionary variation in RNAs that control gene regulation than is observed in regulatory proteins.

RNAi-related processes could also be important for surviving periods of stress. For example, if an organism's environment changes radically, extensive alterations might occur in the transcription of both strands of genomic DNA²⁹. Large segments of the genome could thereby be 'sampled' for short RNA regulators that are advantageous for survival. At present, such events are all conjecture — made possible by the discovery of the many roles of short RNAs in gene regulation and biology.

Carl D. Novina and Phillip A. Sharp are at the Center for Cancer Research, and Phillip A. Sharp is in the Department of Biology and the McGovern Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

e-mails: sharp@mit.edu; cdnovina@mit.edu

1. Fire, A. *et al.* *Nature* **391**, 806–811 (1998).
2. Napoli, C., Lemieux, C. & Jorgensen, R. *Plant Cell* **2**, 279–289 (1990).
3. van der Krol, A. R. *et al.* *Plant Cell* **2**, 291–299 (1990).
4. Wassenege, M. *et al.* *Cell* **76**, 567–576 (1994).
5. Hamilton, A. J. & Baulcombe, D. C. *Science* **286**, 950–952 (1999).
6. Tuschl, T. *et al.* *Genes Dev.* **13**, 3191–3197 (1999).
7. Zamore, P. D. *et al.* *Cell* **101**, 25–33 (2000).
8. Elbashir, S. M. *et al.* *Genes Dev.* **15**, 188–200 (2001).
9. Elbashir, S. M. *et al.* *Nature* **411**, 494–498 (2001).
10. Bernstein, E. *et al.* *Nature* **409**, 363–366 (2001). See also www.nature.com/focus/rnai/animations/animation/animation.htm
11. Matzke, M. *et al.* *Biochim. Biophys. Acta* **1677**, 129–141 (2004).
12. Ketting, R. F. *et al.* *Cell* **99**, 133–141 (1999).
13. Tabara, H. *et al.* *Cell* **99**, 123–132 (1999).
14. Xie, Z. *et al.* *PLoS Biol.* **2**, E104 (2004).
15. Pal-Bhadra, M. *et al.* *Science* **303**, 669–672 (2004).
16. Schramke, V. & Allshire, R. *Science* **301**, 1069–1074 (2003).
17. Lim, L. P. *et al.* *Genes Dev.* **17**, 991–1008 (2003).
18. Lim, L. P. *et al.* *Science* **299**, 1540 (2003).
19. Griffiths-Jones, S. *Nucleic Acids Res.* **32**, D109–D111 (2004).
20. Lee, R. C., Feinbaum, R. L. & Ambros, V. *Cell* **75**, 843–854 (1993).
21. Reinhart, B. J. *et al.* *Nature* **403**, 901–906 (2000).
22. Bartel, D. P. *Cell* **116**, 281–297 (2004).
23. Lewis, B. P. *et al.* *Cell* **115**, 787–798 (2003).
24. Yekta, S. *et al.* *Science* **304**, 594–596 (2004).
25. Doench, J. G. *et al.* *Genes Dev.* **17**, 438–442 (2003).
26. Mochizuki, K. *et al.* *Cell* **110**, 689–699 (2002).
27. Ambros, V. *et al.* *Curr. Biol.* **13**, 807–818 (2003).
28. Kuwabara, T. *et al.* *Cell* **116**, 779–793 (2004).
29. Causton, H. C. *et al.* *Mol. Cell* **12**, 323–337 (2001).
30. Verdel, A. *et al.* *Science* **303**, 672–676 (2004).
31. Lee, Y. *et al.* *Nature* **425**, 415–419 (2003).

Braiding patterns on an inclined plane

The changing boundaries of a stream flowing at a constant rate are explained.

A jet of fluid flowing down a partially wetting, inclined plane usually meanders but — by maintaining a constant flow rate — meandering can be suppressed, leading to the emergence of a beautiful braided structure. Here we show that this flow pattern can be explained by the interplay between surface tension, which tends to narrow the jet, and fluid inertia, which drives the jet to widen. These observations dispel misconceptions about the relationship between braiding and meandering that have persisted for over 20 years.

The flow of water down a partially wetting, inclined surface, as in rivers and streams, is affected by the substrate's roughness and by flow disturbances. This results in seemingly random height and width variations and meandering. River morphology is also influenced by soil erosion¹, which is not necessarily present in the general case of flow down an inclined surface.

We observe that the meandering of a stream on a smooth, non-eroding, inclined plane is caused entirely by upstream disturbances, and meandering can therefore be eliminated, contrary to established belief². However, the variations in the height and width of a braided stream represent an inherent instability of the stream caused by the interaction of surface tension and inertia.

To visualize the stream dynamics, a flat incline (such as an acrylic plate) can be placed under a faucet: the small fluctuations always present in tap water will cause this stream to meander. In our experiment, the meandering is eliminated by maintaining a constant flow rate; we then see a stationary braiding pattern in which the width of the stream expands and contracts as it propagates (for methods, see supplementary information).

This braiding pattern can be explained as follows. When the fluid jet strikes the inclined surface, it spreads out owing to the inertia of the impact. Most of the fluid flows at the outer boundaries of the flow, and the interior of the stream is very shallow. Surface tension limits the extent of the spreading and pulls the outer boundaries of the flow back together. In the process of contraction, the outer edges accelerate beyond equilibrium and 'bounce' on impact, forcing the boundaries outwards; the outer edges then collapse again because of surface tension, and the process repeats.

The amplitude of the subsequent bounces decreases owing to viscous dissipation, and far downstream the flow assumes a simple profile with a part-circular cross-section, when all the forces are in balance^{3,4}.

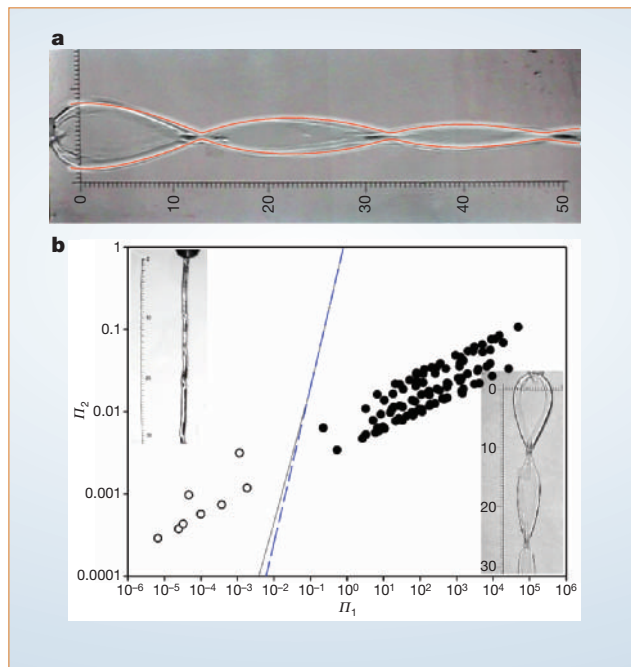


Figure 1 Predicted and experimental braiding patterns in water. **a**, Experimental observation of braiding flow. Parameters varied are flow rate q , inclination angle α and viscosity ν . Here $q = 12.2 \text{ cm}^3 \text{ s}^{-1}$, $\alpha = 45^\circ$, $\nu = 0.016 \text{ cm}^2 \text{ s}^{-1}$. The flow also depends on surface tension, γ , and acceleration due to gravity, g . Red line indicates agreement with theory. Scale in centimetres; flow direction is left to right. **b**, Experimentally observed braiding (filled circles) and non-braiding (open circles) flow represented by the dimensionless parameters Π_1 and Π_2 , where $\Pi_1 = \frac{1}{2} \nu \rho^2 q^5 (g \sin \alpha)^4 \gamma^{-7}$ and $\Pi_2 = \frac{1}{2} \nu \rho^2 q (g \sin \alpha) \gamma^{-2}$. Solid line, theoretical transition boundary from non-braiding to braiding flow; dashed line, power-law fit $\Pi_2 = 1.53 \Pi_1^{0.89}$.

This structure is reminiscent of the fluid chain structure produced by two fluid jets colliding in air⁵, although it is different in its physics because of dissipation at the solid surface.

To explain the braiding phenomenon quantitatively, we constructed a model that assumes that the stream is shallow, the downstream velocity component dominates the flow, and the contact angle between the plate and fluid is constant (see supplementary information). Previous, more complex models^{6,7} do not account for the large amplitude variations that we observe. Our simple model incorporates inertial effects in flows down an inclined plane^{8,9} and couples two ordinary differential equations that predict both the nonlinear evolution of the braids and the transition from the braiding to non-braiding situation.

Although our assumption of a constant contact angle is an approximation¹⁰, it greatly simplifies the analysis. Figure 1 shows that the agreement between the model and experiment is excellent throughout the parameter range investigated; it also shows that the braid-length dependence on the parameters Π_1 and Π_2 (see supplementary information) can account for the transition to a dissipation-dominated, non-braiding flow.

Keith Mertens, Vakhtang Putkaradze, Peter Vorobieff

The University of New Mexico, Albuquerque, New Mexico 87131, USA
e-mail: putkarad@math.unm.edu

- Ikeda, S., Parker, G. & Sawai, K. *J. Fluid Mech.* **112**, 363–377 (1981).
 - Nakagawa, T. & Scott, J. *J. Fluid Mech.* **149**, 89–99 (1984).
 - Davis, S. H. *J. Fluid Mech.* **98**, 225–242 (1980).
 - Perazzo, C. A. & Gratton, J. *J. Fluid Mech.* (in the press).
 - Hasha, A. & Bush, J. W. M. *Phys. Fluids* **14** (suppl.), 8 (2002).
 - Weiland, R. H. & Davis, S. H. *J. Fluid Mech.* **107**, 261–280 (1981).
 - Young, G. W. & Davis, S. H. *J. Fluid Mech.* **176**, 1–31 (1987).
 - Bohr, T., Putkaradze, V. & Watanabe, S. *Phys. Rev. Lett.* **79**, 1038–1041 (1997).
 - Lopez, P. G., Miksis, M. J. & Bankoff, S. G. *Phys. Fluids* **9**, 2177–2183 (1997).
 - DeGennes, P. G. *Rev. Mod. Phys.* **57**, 827–863 (1985).
- Supplementary information accompanies this communication on Nature's website.
Competing financial interests: declared none.

brief communications arising online
www.nature.com/bca

Physiology: Does gut hormone PYY_{3–36} decrease food intake in rodents?

M. Tschöp, T. R. Castañeda, H. G. Joost, C. Thöne-Reineke, S. Ortmann, S. Klaus, M. M. Hagan, P. C. Chandler, K. D. Oswald, S. C. Benoit, R. J. Seeley, K. P. Kinzig, T. H. Moran, A. G. Beck-Sickingler, N. Koglin, R. J. Rodgers, J. E. Blundell, Y. Ishii, A. H. Beattie, P. Holch, D. B. Allison, K. Raun, K. Madsen, B. S. Wulff, C. E. Stidsen, M. Birringer, O. J. Kreuzer, M. Schindler, K. Arndt, K. Rudolf, M. Mark, X. Y. Deng, D. C. Withcomb, H. Halem, J. Taylor, J. Dong, R. Datta, M. Culler, S. Craney, D. Flora, D. Smiley, M. L. Heiman (doi:10.1038/nature02665)

Reply: R. L. Batterham, M. A. Cowley, C. J. Small, H. Herzog, M. A. Cohen, C. L. Dakin, A. M. Wren, A. E. Brynes, M. J. Low, M. A. Ghatei, R. D. Cone, S. R. Bloom (doi:10.1038/nature02666)

Physiology

Does gut hormone PYY₃₋₃₆ decrease food intake in rodents?

Arising from: R. L. Batterham *et al.* *Nature* **418**, 650–654 (2002)

Batterham *et al.* report that the gut peptide hormone PYY₃₋₃₆ decreases food intake and body-weight gain in rodents¹, a discovery that has been heralded as potentially offering a new therapy for obesity. However, we have been unable to replicate their results. Although the reasons for this discrepancy remain undetermined, an effective anti-obesity drug ultimately must produce its effects across a range of situations. The fact that the findings of Batterham *et al.*¹ cannot easily be replicated calls into question the potential value of an anti-obesity approach that is based on administration of PYY₃₋₃₆.

We have been unable to observe an acute or chronic decrease in food intake or in body weight following peripheral administration of PYY₃₋₃₆ to rodents (for full details, see the authors' website at www.pyyobesity.com). We repeated and extended the original studies¹, peripherally and/or centrally administering PYY₃₋₃₆ in equivalent concentrations to rodents from eight different strains in twelve independent laboratories (our methods and results are summarized in Table 1).

Food intake was measured continuously or hourly for up to 24 hours and daily for up to 10 days in 'acute' and 'chronic' experiments. In 37 of 39 experiments, PYY₃₋₃₆ either did not change or increased food intake compared with control-treated rodents (Table 1; see also Fig. 1a). In two continuous-infusion experiments, food intake did decrease, but only transiently; no effects were found on meal size or number, although we observed a transient increase in feeding behaviour that lasted for 10 min.

We found that the anorexigenic agents MT-II (a melanocortin-receptor agonist) and naloxone (an opioid-receptor antagonist) decreased food intake ($P=0.001$) in rats that had not responded to PYY₃₋₃₆, as did the gut-satiety hormone cholecystokinin ($P=0.05$) (see Fig. 1b, c and the authors' website). This indicated that the animals and conditions that we used were capable of revealing known anorectic effects.

Body weight was measured for up to 10 days following peripheral treatment with PYY₃₋₃₆ (Table 1; see also Fig. 1d). In 13 of 14 experiments, body weight either did not change or increased compared with that of control-treated rodents; only in one case (*Ob/Ob*^{-/-} mouse) was there a transient decrease in body weight (Table 1). Lean-muscle mass did not change, and adipose-tissue mass either increased or did not change.

We found that there was a significant reduction in the animals' locomotor activity and rearing and sniffing behaviour following PYY₃₋₃₆ administration; overall energy

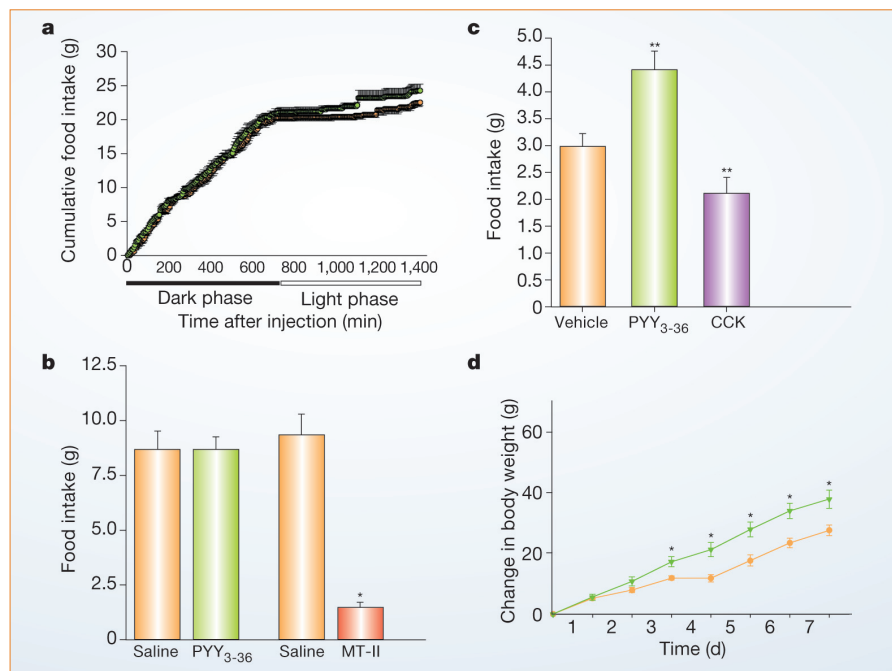


Figure 1 Lack of inhibitory effect of PYY₃₋₃₆ on food intake in rodents. Animals were acclimated to test conditions for 5–28 days (Table 1), fed normally and showed no signs of stress. **a–d**, Testing of human PYY₃₋₃₆ (**a**, **b**, **d**) and rat PYY₃₋₃₆ (**c**). **a**, PYY₃₋₃₆ (intraperitoneal (i.p.) dose of 3 $\mu\text{g kg}^{-1}$, $n=10$; green circles) did not decrease food intake over 24 hours in individually housed Wistar rats compared with vehicle-injected controls (orange circles). **b**, The melanocortin-receptor agonist MT-II (3 mg kg^{-1}), but not PYY₃₋₃₆ (i.p., injected just before lights turned off; 100 $\mu\text{g kg}^{-1}$), decreased 4-hour food intake in Wistar rats ($n=18$). **c**, PYY₃₋₃₆ (i.p., 5 $\mu\text{g kg}^{-1}$) increased food intake after overnight fasting (30 min, $P<0.05$, $n=10$) in Sprague–Dawley rats; cholecystokinin (CCK; i.p., 6 $\mu\text{g kg}^{-1}$) decreased food intake ($n=14$). **d**, PYY₃₋₃₆ (i.p., twice a day at 50 $\mu\text{g kg}^{-1}$; green line) increased body weight compared with saline control (orange line) over a 7-day treatment period in Sprague–Dawley rats ($n=8$). Error bars show standard error of the mean; * $P<0.001$; ** $P<0.05$.

expenditure tended to decrease (not significant; 6 mice received PYY₃₋₃₆), as did faecal caloric content (not significant; 8 mice received PYY₃₋₃₆). These changes favour a positive energy balance and ultimately promote body-fat gain. Experiments usually had equal or greater statistical power (number of animals in a group) compared with those of Batterham *et al.*¹.

Batterham *et al.* do not describe in detail the adaptation or habituation of their animals¹, so we used our own established and more extensive adaptation protocols. We carefully adapted our animals to experimental conditions for at least five days and excluded stress-related bias in these animals by standard assay of plasma corticosteroids and hypothalamic *c-fos* expression in the paraventricular nucleus area, or by monitoring feeding² and other stress-sensitive behaviour (results not shown).

Batterham *et al.* used human PYY₃₋₃₆ (supplied by Bachem) for their rodent studies (personal communications); we ran our tests with human or rodent PYY₃₋₃₆ from several suppliers (Bachem, P&E GmbH, Polypeptide Laboratories, Neosystems, Eli Lilly). The relative molecular mass of the peptide was confirmed upon delivery and in the injection

solution either before or after experiments (by high-performance liquid chromatography and/or mass spectrometry).

The biological activity of PYY₃₋₃₆ was confirmed *in vivo* by determining the increase in food intake following intracerebroventricular administration (through Y1/Y5 receptors³) and the delay in gastric emptying following intraperitoneal administration (through Y2 receptors⁴). These assays test the functional integrity of the peptide and show that its activity is not species specific. Y2-receptor binding was assayed⁵ in addition to confirm the activity and integrity of the fresh and left-over PYY₃₋₃₆.

To our knowledge, only one group not associated with the authors of refs 1 and 6 has so far replicated parts of the original study. However, this group^{7,8} found no effect of PYY₃₋₃₆ on body weight or on chronic food intake, and no effect on acute food intake without specific pre-fasting. Others have studied the longer peptide PYY₁₋₃₆ (ref. 9) or focused on the cardiovascular effects of PYY₃₋₃₆ (ref. 10). On the basis of the results available so far in rodents, we believe that there is no conclusive evidence that PYY₃₋₃₆ is a physiological satiety factor with the potential to treat obesity.

Table 1 **Effects of PPY₃₋₃₆ on food intake and body weight in rodents**

Species (male)	Administration route	Study	Dose	Total number of animals (controls)*	Food intake	Body weight
Wistar rat	i.c.v.	Acute	10 µg 5 µl ⁻¹	14 (5)	Increase (~200%)	ND
			2 µg 3 µl ⁻¹ , 4 µg 3 µl ⁻¹ , 8 µg 3 µl ⁻¹	32 (8)	Increase (~300%)	ND
	i.p.	Acute	0.3 µg 100 g ⁻¹	10 (5)	No effect	ND
			100 µg kg ⁻¹	16 (8)	No effect	ND
			100 µg kg ⁻¹	24 (12)	No effect	ND
			100 µg kg ⁻¹	24 (8)	No effect	ND
			3 µg, 30 µg, 100 µg (twice a day)	32 (8)	No effect	ND
			100 µg kg ⁻¹	24 (6)	No effect	ND
			5 µg kg ⁻¹	102 (32)	Increase	ND
			30 µg kg ⁻¹ , 100 µg kg ⁻¹ , 300 µg kg ⁻¹	32 (8)	No effect	ND
		Chronic (2 d)				
Sprague-Dawley rat	i.v.	Chronic (2 d)	5 µg 100 g ⁻¹ (twice a day)	16 (8)	No effect	No effect
	i.p.	Acute	3 µg 100 g ⁻¹ , 10 µg 100 g ⁻¹	72 (24)	No effect	No effect
			3 µg 100 g ⁻¹	96 (24)	No effect	ND
			0.3 nmol kg ⁻¹ , 1 nmol kg ⁻¹ , 3 nmol kg ⁻¹ , 10 nmol kg ⁻¹	5 × 8 (8)	Transient decrease†	ND
	Chronic (7 d)		5 µg 100 g ⁻¹ (twice a day)	16 (8)	Increase (~7%)	Increase (~35%)
Long Evans rat	i.c.v.	Acute	0.2 nmol, 1 nmol	24 (8)	No effect‡, increase§	No effect
	i.p.	Acute	30 µg kg ⁻¹ , 100 µg kg ⁻¹ , 300 µg kg ⁻¹	32 (8)	No effect	No effect
			30 µg kg ⁻¹	10 (5)	No effect	No effect
			5 µg 100 g ⁻¹	16 (8)	ND	No effect
	Chronic (5 d)					
Lister hooded rat	i.p.	Acute	3 µg kg ⁻¹ , 30 µg kg ⁻¹ , 100 µg kg ⁻¹	4 × 10 (10)	No effect	No effect
NMRI mouse	i.p.	Acute	50 µg kg ⁻¹ , 100 µg kg ⁻¹ (twice a day)	18 (6)	No effect	No effect
		Chronic (10 d)	100 µg kg ⁻¹ (twice a day)	14 (7)	Increase (~20%)	Increase (~5%)
C57BL/6 mouse	s.c.	Chronic (7 d)	1 mg kg ⁻¹	17 (9)	Transient decrease¶	No effect
	i.p.	Acute	1 nmol kg ⁻¹ , 3 nmol kg ⁻¹	3 × 6 (6)	No effect	ND
			1 nmol kg ⁻¹ , 3 nmol kg ⁻¹	4 × 6 (12)	No effect	ND
Ob/Ob ^{-/-} mouse	s.c.	Chronic (7 d)	1 mg kg ⁻¹	17 (9)	Transient decrease¶	Transient decrease#
NZO mouse	i.p.	Chronic (8 d)	100 µg kg ⁻¹	10 (4)	No effect	No effect

The results were obtained after habituation of all animals, which were individually housed. Briefly, animals were adapted to single-housed test conditions and diet, non-interrupted 12-hour light cycles and handling, including body-weight measurements and saline injections for 5–28 days before study onset. Food intake in the pre-study habituation phase was monitored to verify that it had stabilized and did not differ from the food intake of the control group during testing. Bedding was screened for spilt food and noise was kept to a minimum. Animals were monitored for evidence of stress (for example, poor grooming, defensive postures, discoloration around eyes and nose). Although the results of a few experiments were in agreement with those of Batterham *et al.*, that number (assuming all tests performed at a 2-tailed 0.05 significance level) was not significantly greater than expected by chance ($P = 0.255$ for food intake), whereas the number in the other direction was significantly greater than would be expected by chance ($P = 0.016$). For further details, see the authors' website at www.pyobesity.com.

ND, not determined; i.c.v., intracerebroventricular administration; i.p., intraperitoneal administration; s.c., continuous administration by subcutaneously implanted osmotic minipump.

* The total number of animals comprises treated animals and controls.

† Transient decrease (about 70%) at 30 min only at the dose 1 nmol kg⁻¹ only. No effect with any other dose or at any other time point.

‡ No effect at 0.2 nmol dose.

§ Increase (about 60%) at 1 nmol dose.

|| No effect on total 1-h intake, but significant increase (about 23–50%) in duration of feeding behaviour in the first 10 min of test.

¶ Transient decrease (about 30%) in first 3 d, then no effect.

Decrease (about 2%) on first day, then no effect.

M. Tschöp^{1,2}, T. R. Castañeda^{1,2}, H. G. Joost², C. Thöne-Reineke³, S. Ortmann^{2,4}, S. Klaus², M. M. Hagan⁵, P. C. Chandler⁵, K. D. Oswald⁵, S. C. Benoit¹, R. J. Seeley¹, K. P. Kinzig⁶, T. H. Moran⁶, A. G. Beck-Sickinger⁷, N. Koglin⁷, R. J. Rodgers⁸, J. E. Blundell⁸, Y. Ishii⁸, A. H. Beattie⁸, P. Holch⁸, D. B. Allison⁹, K. Raun¹⁰, K. Madsen¹⁰, B. S. Wulff¹⁰, C. E. Stidsen¹⁰, M. Birringer¹¹, O. J. Kreuzer¹¹, M. Schindler¹², K. Arndt¹², K. Rudolf¹², M. Mark¹², X. Y. Deng¹³, D. C. Withcomb¹³, H. Halem¹⁴, J. Taylor¹⁴, J. Dong¹⁴, R. Datta¹⁴, M. Culler¹⁴, S. Craney¹⁵, D. Flora¹⁵, D. Smiley¹⁵, M. L. Heiman¹⁵

¹Department of Psychiatry, University of Cincinnati Genome Research Institute, Cincinnati, Ohio 45237, USA

Ohio 45237, USA
e-mail: tschoemh@ucmail.uc.edu

²German Institute of Human Nutrition, Potsdam-Rehbrücke 14558, Germany; ³Center for Cardiovascular Research, Institute for Pharmacology and Toxicology, Charité University Hospital, Berlin 10115, Germany; ⁴Institute for Zoo and Wildlife Research, Berlin 10315, Germany; ⁵Department of Psychology, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA; ⁶Department of Psychiatry and Behavioral Science, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA; ⁷Institute of Biochemistry, University of Leipzig, Leipzig 04103, Germany; ⁸Behavioural Pharmacology Laboratory, School of Psychology, University of Leeds, Leeds LS2 9JT, UK; ⁹Section on

Statistical Genetics, Department of Biostatistics and Clinical Nutrition Research Center, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA; ¹⁰Novo Nordisk A/S, Discovery, Maalov 2760, Denmark; ¹¹Peptides and Elephants GmbH, Potsdam-Rehbrücke 14558, Germany; ¹²Boehringer-Ingelheim Pharma GmbH and Co. KG, Biberach 88397, Germany; ¹³Division of Gastroenterology, Hepatology and Nutrition, Department of Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15213, USA; ¹⁴Biomeasure Incorporated/Ipsen Group, Milford, Massachusetts 01757, US; ¹⁵Eli Lilly & Co. Research Laboratories, Indianapolis, Indiana 46285, USA
doi:10.1038/nature02665

1. Batterham, R. L. *et al.* *Nature* **418**, 650–654 (2002).

2. Halford J. C. *et al.* *Pharmacol. Biochem. Behav.* **61**,

- 159–168 (1998).
- Kanatani, A. *et al. Endocrinology* **141**, 1011–1016 (2000).
 - Allen, J. M. *et al. Digestion* **30**, 255–262 (1984).
 - Amersham Biosciences, RPNQ0085, Neuropeptide Y SPA binding assay (Y2).
 - Halatchev, I. G. *et al. Endocrinology* **145**, 2585–2590 (2004).
 - Challis, B. G. *et al. Biochem. Biophys. Res. Commun.* **311**, 915–919 (2003).
 - Challis, B. G. *et al. Proc. Natl Acad. Sci. USA* **101**, 4695–4700 (2004).
 - Asakawa, A. *et al. Gastroenterology* **124**, 1325–1336 (2003).
 - Nordheim, U. & Hofbauer, K. G. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **286**, R138–R142 (2004).

Competing financial interests: declared. (Several of the authors consult with, and have received financial support from, companies and institutions, including pharmaceutical firms, that may have active programmes relating to PYY_{3–36}, of which we are not aware. Several of these companies have obesity-treatment development programmes and may therefore, from some perspectives, be considered competitors to companies pursuing PYY_{3–36}.)

Batterham et al. reply — The results of Tschöp *et al.*¹ on the lack of effect of peripheral administration of PYY_{3–36} on food intake in rodents are at odds with both the published literature^{2–8} (our Table 1) and with earlier data generated by the Tschöp laboratory.

It has been shown that PYY_{3–36} does not inhibit food intake in animals whose appetite is reduced by stress⁵ and this seems the likeliest explanation for the different findings now reported by Tschöp *et al.* The results of Tschöp *et al.* may also be undermined by their failure to measure the blood concentrations of PYY_{3–36} that were achieved following peripheral administration. Were they sufficient to result in inhibition of food

intake? In addition, their meta-analysis does not include earlier data generated in Tschöp's laboratory that confirm our findings⁹ (M.T., personal communication; data available from S.R.B. on request).

Tschöp *et al.* do not challenge the large body of appropriate mechanistic data that includes the inhibitory effect on food intake of intra-arcuate Y2-receptor (Y2R) agonists or PYY_{3–36}, changes in hypothalamic neuropeptide messenger RNA, the analysis of rodent hypothalamic explants and studies in Y2R-knockout mice^{4,9}. The human studies showing that PYY_{3–36} inhibits food intake^{9,10} are also uncontested.

Peripherally administered PYY_{3–36} inhibits food intake in procedure-acclimatized

Table 1 Summary of literature data on PYY_{3–36} and food intake in rodents

Species	Route	Study	Dose	Total number of animals (controls)	Effect
Wistar rats (male)	i.p.	Acute (freely feeding) ⁶	300 µg kg ⁻¹	28 (14)	Reduction in 24-h food intake*
	i.p.	Acute (dark phase) ⁹	3, 10, 30 µg kg ⁻¹	32 (8)	Reduction in 4-h food intake*
	i.p.	Acute (fasted) ⁹	3, 10, 30 µg kg ⁻¹	32 (8)	Reduction in 4-h food intake*
	i.p.	Acute (fasted) ⁹	50 µg kg ⁻¹	24 (12)	Reduction in 24-h food intake**
	i.p.	Chronic (7 d) ⁹	50 µg kg ⁻¹ d ⁻¹ (twice daily)	24 (12)	Reduction in food intake and body-weight gain**
	Intra-arcuate	Acute (fasted) ⁹	0.001, 0.01, 0.1 nmol	32 (8)	Reduction in 2-h food intake*
Diabetic Fatty Zucker rats (male)	Alzet pump	Chronic (28 d) ⁷	30, 100, 300 µg d ⁻¹	44 (14)	Reduction in cumulative food intake at highest two doses** Reduction in HbA1c and fructosamine*
Fatty Zucker obese non-diabetic rats (male)	Alzet pump	Chronic (56 d) ⁷	100 µg kg ⁻¹ d ⁻¹	20 (13)	Reduction in cumulative food intake and body weight*
129/J mice (male)	i.p.	Acute (fasted) ⁴	100 µg kg ⁻¹	16 (8)	Food intake reduction at 6 h*** and 24 h**
	i.p.	Acute (dark phase) ⁴	100 µg kg ⁻¹	16 (8)	Tendency to reduce 24-h food intake (NS)
	i.p.	Acute (freely feeding) ⁴	100 µg kg ⁻¹	20 (10)	Increased POMC mRNA at 6 h** and 24 h** Decreased NPY mRNA at 6 h**
POMC KO mice (male) and WT controls (male)	i.p.	Acute (fasted) ³	100 µg kg ⁻¹	64 (32)	Reduction in 4-h food intake**
	i.p.	Chronic (7 d) ³	100 µg kg ⁻¹ (twice daily)	32 (16)	No effect on food intake or body weight
NIH/Swiss non-obese mice (female)	i.p.	Acute (fasted) ⁷	0.1, 1, 3, 10, 30, 100, 300, 1,000 µg kg ⁻¹	192 (40)	Reduction in 1-h food intake with all doses greater than 3 µg kg ⁻¹ *
Ob/Ob ^{-/-} mice (female)	Alzet pump	Chronic (28 d) ⁷	100, 300, 1,000 µg kg ⁻¹ d ⁻¹	32 (8)	300 and 1,000 µg kg ⁻¹ d ⁻¹ doses led to reduction in cumulative food intake** and reduction in body-weight gain***
C57BL6/J DIO mice (male)	Alzet pump	Chronic (28 d) ⁷	30, 100, 300, 1,000 µg kg ⁻¹	76 (20)	Dose-dependent reduction in cumulative food intake*, body weight*** and reduced adiposity*
C57BL6/J mice (male) ⁵	i.p.	Acute (fasted) unacclimatized ⁵	3, 30, 100 µg kg ⁻¹	20 (5)	No effect on food intake in unacclimatized mice
	i.p.	Acute (fasted) ⁵	3, 30, 100 µg kg ⁻¹	20 (5)	Reduction in 4-h food intake*
	i.p.	Acute (fasted) ⁵	3, 100 µg kg ⁻¹	19 (10)	Reduction in 4-h food intake*
	i.p.	Acute (fasted) ⁵	30, 100 µg kg ⁻¹	19 (10)	Reduction in 4-h food intake*
Y2 receptor KO mice (males) and WT controls (males)	i.p.	Acute (fasted) ⁹	50 µg kg ⁻¹	40 (10)	Reduction of food intake in WT mice * No effect in Y2R KO mice
Ddy mice (male)	i.p.	Acute (fasted) ²	3 nmol	18 (9)	Reduction in 24-h food intake*
	i.p.	Acute (dark phase) ²	3 nmol	16 (8)	Reduction in 24-h food intake*
MC4R KO mice (male) and WT controls (male)	i.p.	Acute (dark phase) ⁵	3, 30, 100 µg kg ⁻¹	70 (36)	Reduction in 4-h food intake
Agouti mice (male) and WT controls (male)	i.p.	Acute (fasted) ⁵	0.02 µmol kg ⁻¹	44 (21)	Reduction in 2-h re-feeding***

Abbreviations: KO, knockout; NS, non-significant; i.p., intraperitoneal; WT, wild type; POMC, pro-opiomelanocortin; MC4R, melanocortin-4-receptor. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

Corrigendum

Physiology: Does gut hormone PYY₃₋₃₆ decrease food intake in rodents?

M. Tschöp, T. R. Castañeda, H. G. Joost,
C. Thöne-Reineke, S. Ortmann, S. Klaus, M. M. Hagan,
P. C. Chandler, K. D. Oswald, S. C. Benoit, R. J. Seeley,
K. P. Kinzig, T. H. Moran, A. G. Beck-Sickinger, N. Koglin,
R. J. Rodgers, J. E. Blundell, Y. Ishii, A. H. Beattie,
P. Holch, D. B. Allison, K. Raun, K. Madsen, B. S. Wulff,
C. E. Stidsen, M. Birringer, O. J. Kreuzer, M. Schindler,
K. Arndt, K. Rudolf, M. Mark, X. Y. Deng, D. C. Withcomb,
H. Halem, J. Taylor, J. Dong, R. Datta, M. Culler,
S. Craney, D. Flora, D. Smiley, M. L. Heiman
(doi:10.1038/nature02665)

The name of one author (D. C. Whitcomb) was misspelled in this Brief Communications Arising.

Corrigendum doi:10.1038/nature03019

rodents⁹, and these findings have since been supported by others^{2–8}. Published studies confirming this effect (see our Table 1) are more extensive than the data presented by Tschöp *et al.* in their Table 1. PYY_{3–36} reproducibly inhibits food intake in mice after an overnight fast and during dark-phase feeding in a dose-dependent manner^{2–8}, but food intake in rodents is also inhibited by stress⁵.

Although Tschöp *et al.* claim that they performed assays to exclude stress, how can these be done in an ongoing feeding study? For example, the assessment of *c-fos* by immunohistochemistry necessitates killing the animal. Moreover, Tschöp *et al.* did demonstrate anorectic effects of PYY_{3–36} when using non-stressful methods, such as infusion of the peptide with a minipump. They show in their Table 1 that food intake was reduced for three days in both C57BL/6 and *Ob/Ob*^{−/−} mice when using this non-stressful route of administration.

Contrary to their claims to have repeated our original studies, Tschöp *et al.* do not present data examining the effects of administering PYY_{3–36} to mice given food after fasting. After feeding, endogenous PYY_{3–36} is already raised and administration of exogenous PYY_{3–36} is not effective. Many of the para-

digms listed in their Table 1 are not relevant because the administration of PYY_{3–36} was either intracerebroventricular (when PYY_{3–36} stimulates food intake through the NPY Y1/5 receptor) or to satiated animals.

We question the results presented in Fig. 1c of Tschöp *et al.* The data seem to be derived from two separate experiments that have been recombined. In those two experiments, there is no direct comparison between PYY_{3–36} and CCK — the first compares CCK with PYY_{1–36} and the second compares PYY_{3–36} with other peptides (see their website www.pyyobesity.com).

Tschöp *et al.* question the potential of PYY_{3–36} as a treatment for obesity. However, Pittner *et al.*⁷ have demonstrated that PYY_{3–36} reduces food intake, decreases adiposity and improves glycaemic indices in several rodent models of obesity (see our Table 1).

This exchange exemplifies the fact that feeding is a behaviour and is thus very sensitive to the environment in which it is studied. Stress-inhibited food intake is a particularly poor circumstance in which to try to study another feeding inhibitor.

R. L. Batterham, M. A. Cowley, C. J. Small, H. Herzog, M. A. Cohen, C. L. Dakin, A. M. Wren, A. E. Brynes, M. J. Low,

M. A. Ghatei, R. D. Cone, S. R. Bloom*

*Imperial College Faculty of Medicine at Hammersmith Campus, London W12 0NN, UK
e-mail: s.bloom@imperial.ac.uk
doi:10.1038/nature02666

1. Tschöp, M. *et al.* *Nature* doi:10.1038/nature02665 (2004).
 2. Asakawa, A. *et al.* *Gut* **52**, 947–952 (2003).
 3. Challis, B. G. *et al.* *Proc. Natl Acad. Sci. USA* **101**, 4695–4700 (2004).
 4. Challis, B. G. *et al.* *Biochem. Biophys. Res. Commun.* **311**, 915–919 (2003).
 5. Halatchev, I. G., Ellacott, K. L., Fan, W. & Cone, R. D. *Endocrinology* **145**, 2585–2590 (2004).
 6. Nordheim, U. & Hofbauer, K. G. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **286**, R138–R142 (2004).
 7. Pittner, R. A. *et al.* *Int. J. Obes.* advance online publication, 15 June 2004 (doi:10.1038/sj.ijo.0802696).
 8. Martin, N. M. *et al.* *Int. J. Obes.* advance online publication, 18 May 2004 (doi:10.1038/sj.ijo.0802646).
 9. Batterham, R. L. *et al.* *Nature* **418**, 650–654 (2002).
 10. Batterham, R. L. *et al.* *N. Engl. J. Med.* **349**, 941–948 (2003).
- Competing financial interests:** declared. (R.L.B., M.A.C., C.J.S., M.A.G. and S.R.B. have a financial interest in technology licensed to Thiakis, a company that may have a commercial interest in the results of this research and technology. M.A.C. has a significant financial interest in Orexigen Therapeutics, a company that may have a commercial interest in the results of this research and technology. This potential conflict of interest has been reviewed and managed by the OHSU Conflict of Interest in Research Committee. R.D.C. is a significant shareholder in Neurocrine Biosciences, chief scientific officer of and a significant shareholder in Znomics, and a co-inventor on patents dealing with PYY, licensed to Orexigen and other biotechnology companies. H.H., M.A.C., C.L.D., A.M.W., A.E.B. and M.J.L. have no conflicts of interest.)

Prediction of Emperor-Hawaii seamount locations from a revised model of global plate motion and mantle flow

Bernhard Steinberger^{1*}, Rupert Sutherland² & Richard J. O'Connell³

¹*Institute for Frontier Research on Earth Evolution (IFREE), Japan Agency for Marine-Earth Science and Technology (JAMSTEC) (previously Japan Marine Science and Technology Center), 2-15 Natsushima-cho, Yokosuka 237-0061, Japan*

²*Institute of Geological and Nuclear Sciences, PO Box 30-368, Lower Hutt, New Zealand*

³*Department of Earth and Planetary Sciences, Harvard University, 20 Oxford Street, Cambridge, Massachusetts 02138, USA*

* Present address: Bayerisches Geoinstitut, Universität Bayreuth, D-95440 Bayreuth, Germany. After 1 August 2004: Norges geologiske undersøkelse, N-7491 Trondheim, Norway

The bend in the Hawaiian-Emperor seamount chain is a prominent feature usually attributed to a change in Pacific plate motion ~47 Myr ago. However, global plate motion reconstructions fail to predict the bend. Here we show how the geometry of the Hawaiian-Emperor chain and other hotspot tracks can be explained when we combine global plate motions with intraplate deformation and movement of hotspot plumes through distortion by global mantle flow. Global mantle flow models predict a southward motion of the Hawaiian hotspot. This, in combination with a plate motion reconstruction connecting Pacific and African plates through Antarctica, predicts the Hawaiian track correctly since the date of the bend, but predicts the chain to be too far west before it. But if a reconstruction through Australia and Lord Howe rise is used instead, the track is predicted correctly back to 65 Myr ago, including the bend. The difference between the two predictions indicates the effect of intraplate deformation not yet recognized or else not recorded on the ocean floor. The remaining misfit before 65 Myr ago can be attributed to additional intraplate deformation of similar magnitude.

Hotspots are frequently assumed to be fixed and used as a reference frame for plate motions¹. However, when the Hawaiian hotspot track is predicted from a global plate motion chain based on relative plate motion data, and it is assumed that the Hawaiian hotspot is fixed relative to African hotspots, it does not fit the observed track: the predicted track is essentially straight, lies south of the Hawaiian seamount chain, and has no feature corresponding to the Hawaiian-Emperor bend². Explanation of this misfit requires one of the following, or a combination of them, to be true: (1) motion of Pacific plate hotspots, in particular more than 1,000 km southward motion for the Hawaiian hotspot between ~80 Myr ago and the time of the bend and minor southward motion after this time, (2) motion of hotspots in the African hemisphere, (3) motion at an additional plate boundary, or (4) deformation at a diffuse intraplate boundary not included in the plate motion chain. Palaeomagnetic data from the Emperor chain³ yield a southward component of Hawaiian hotspot motion more than 1,000 km relative to the palaeomagnetic axis between ~80 and 49 Myr ago. True polar wander⁴ can explain at most a small fraction of this motion. Hence the southward component of hotspot motion required by palaeomagnetic data is roughly the same as the southward component required to remove the misfit between the observed Hawaiian hotspot track and the track predicted by the plate motion chain². However, to remove the misfit before the bend by solely invoking Pacific plate hotspot motion additionally requires a westward component of Hawaiian hotspot motion; to explain the sharp bend would further require that this motion slowed down abruptly at the time of the bend.

Here we use a model of plumes moved and distorted by global mantle flow to compute hotspot motion, and combine it with plate motions to predict hotspot tracks globally. We show that computed hotspot motion is sufficient to remove any misfit between predicted and observed tracks after the time of the Hawaiian-Emperor bend. Initial radiometric dating of seamounts near the bend yielded ages of ~43 Myr⁵, but recent work suggests that it might be as old as

50 Myr⁶. For the track before the time of the bend we use two different plate motion chains connecting Pacific and African plates. With chain model 1, which goes through Antarctica, the Hawaiian track before the bend, as well as the bend itself, is not correctly predicted. None of our plume motion models yield the westward component of motion of the Hawaiian hotspot that is required for a correct prediction. However, with chain model 2, which goes through Australia and the Lord Howe rise, the bend and the Hawaiian hotspot track for times between 65 Myr ago and the present is predicted.

We restrict our discussion to four hotspots: Hawaii and Louisville in the Pacific hemisphere, and Tristan and Réunion in the African hemisphere. These are both necessary and sufficient for the purpose of this paper. They are necessary because at least two hotspots on each plate are needed to separately determine plate motions relative to hotspots in either hemisphere. They are sufficient because we show that it is not possible to fit these four tracks simultaneously with plate motion chain model 1 at times before the Hawaiian-Emperor bend. Moreover, we have shown previously⁷ that, separately for each hemisphere, a greater number of hotspot tracks can be fitted by our model results that predict hotspot motion. On the basis of circumstantial evidence, these four hotspots are among the seven most likely candidates for a deep mantle origin⁸—an underlying assumption of our model. The three others are probably not suitable for the purpose of this paper: the Easter plume is important for the question of hotspot motion during the past ~47 Myr but not so important for the time before that⁹; the Iceland plume might not have left a clear 'hotspot track'—instead, volcanic features in the region might be caused by plume–ridge interaction¹⁰; and the Afar hotspot probably initiated less than 47 Myr ago^{11–14}.

Hotspot motions predicted by mantle flow models

We compute hotspot motion with a method that was developed and described in detail previously¹⁵. Results shown here are for one

particular model; however, comparison with our previous work^{7,9,15–17} shows that these are typical results. By referring to these previous results, we show that the conclusions here do not depend on the choice of one particular set of model parameters but are robust conclusions that are valid as long as our model is at least

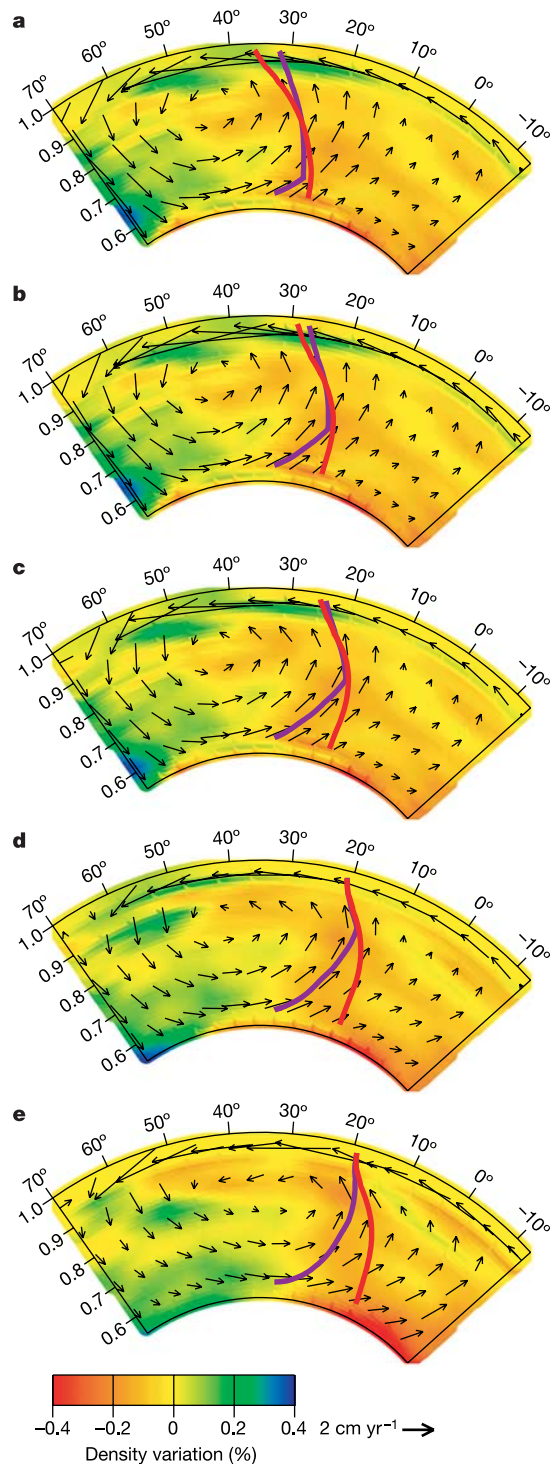


Figure 1 North-south mantle cross-section at 155° W. For different times, computed mantle density anomalies are shown in rainbow colours, and north-south and vertical components of computed mantle flow as arrows. **a**, 120 Myr ago; **b**, 90 Myr ago; **c**, 60 Myr ago; **d**, 30 Myr ago; **e**, present. Also shown is the projection of the predicted Hawaiian plume conduit for source moving with the flow (plume initiation age 170 Myr ago; thick red line) and fixed source (age 150 Myr ago; thick violet line) at those times.

qualitatively correct. The computation consists of two steps. First, a large-scale mantle flow field is computed, which is based on a mantle density structure inferred from seismic tomography, global plate motions and a radial mantle viscosity structure; this model fits several observational constraints (see Supplementary Information 1). The computed large-scale flow field is time dependent, although it does not change much over the period considered here. In the lower part of the mantle, the computed flow field is dominated by structure of spherical harmonic degree two, with large upwellings under the Pacific Ocean and Africa, downwellings around the Pacific Ocean, and flow towards the large upwellings in the lowermost mantle. In the upper part of the mantle, flow is also related to plate motions. In the uppermost part of the lower mantle, computed flow is a combination of outward flow away from the large upwellings and plate return flow towards ridges. The latter might be an important contribution close to spreading ridges. Figure 1 shows a cross-section through density and flow field for different times beneath the Pacific Ocean.

In the second step, a plume conduit is inserted into the flow field. It is taken to be initially vertical. Plume initiation times are assumed to be the same as in ref. 7, except for Hawaii, where the age is unknown and 170 or 150 Myr ago is used here. The base of the plume (assumed at depth 2,620 km, at the top of a low-viscosity layer) is assumed either to move with the horizontal component of flow or to be fixed in location (see Supplementary Information 1). The velocity of points along the plume conduit is computed as the vector sum of ambient mantle flow and a buoyant rising velocity (see Supplementary Information 1). It is assumed that plume conduits do not influence larger-scale mantle flow. Hotspot surface motion is computed from the positions where, over time, the plume conduit reaches the base of the lithosphere (assumed at depth 100 km). Our choice of modelling parameters and assumptions has been explained in previous work^{7,9,15,16}.

Motion of plume conduits in a high-viscosity lower mantle is dominated by advection, and in a low-viscosity upper mantle it is dominated by buoyant rising. Consequently, hotspot motion tends to be similar to the horizontal flow component at the depth at which the transition from low to high viscosity occurs (that is, the upper part of the lower mantle)¹⁶. For the viscosity structure used, the horizontal flow at this depth has a root-mean-square value of $\sim 1 \text{ cm yr}^{-1}$. Beneath the Tristan and Réunion hotspots it is dominated by outward flow from the large upwelling under Africa; in the case of Louisville it is a combination of flow away from the large upwelling under the Pacific Ocean and plate return flow towards the Pacific–Antarctic ridge¹⁷. The computed motion of these hotspots is indeed slow and is as expected from the flow pattern (Fig. 2). For the Hawaiian plume, another effect leads to faster (a few cm yr^{-1}) hotspot motion. Figure 1 shows a projection of the Hawaiian conduit; it is strongly tilted in a north–south direction. Flow in the upper part of the mantle is to the north and in the lower part to the south, which tilts and distorts the plume. Subsequently, the rising of a tilted plume conduit, aided by a large-scale upwelling, causes comparatively rapid hotspot motion. Among the hotspots considered here, we find this effect only for the Hawaiian hotspot.

We have previously shown results for a larger number of models, and discussed the dependence of results on various model parameters^{7,9,15–17}. In summary, these results yield the following. For Hawaii, the motion is in a southerly to southeasterly direction. The average speed is moderately slow ($\sim 1 \text{ cm yr}^{-1}$), but episodes of faster motion (several cm yr^{-1}), lasting for several tens of Myr occur for some models. Computed hotspot motion tends to be faster, if an older age for the plume is assumed. For Louisville there is slow motion ($\sim 1 \text{ cm yr}^{-1}$ or less) in various directions—in most cases in a southeasterly direction. For Réunion there is slow motion ($\sim 1 \text{ cm yr}^{-1}$ or less) in an easterly direction. For Tristan there is slow motion (less than 1 cm yr^{-1}) in a southeasterly to southwesterly direction, or an almost stationary hotspot.

Obviously, it is also possible to construct models with substantially larger hotspot motion (for example, with lower viscosity, and/or larger density anomalies in the mantle¹⁵); however, such models are not consistent with observations, such as hotspot tracks, whereas the models just summarized are broadly consistent. The directions of hotspot motion are less sensitive to model differences than are the speeds of motion.

Relative plate motions and hotspot tracks

We now account for predicted hotspot motion and compute hotspot tracks from specified relative plate motions. We cannot give formal uncertainties, although—on the basis of the spread of model results as just discussed—we can judge whether a discrepancy seems to be significant. The global tectonics of the Earth is characterized by a group of plates that diverge from the Indian and Atlantic oceans, and a Pacific group of oceanic plates that converge around most of

the Pacific rim. The relative motions of the two groups can be determined from observations of the sea floor in the South Pacific and southern Indian Ocean, with the caveat that deformation is known to have occurred in continental regions of Antarctica and New Zealand, but this is hard to quantify.

For times younger than chron 20 (43 Myr ago), substantial Australia–Pacific plate motion (through New Zealand) cannot be determined with sufficient precision from local data for it to be useful in this analysis. However, seafloor spreading in the South Pacific is accurately quantified¹⁸, and motion between East and West Antarctica has been determined from a consideration of seafloor geometry¹⁹. The motion of Africa, relative to East Antarctica since the Late Cretaceous, is determined from sea floor in the Southwest Indian Ocean²⁰. Hence the motion of Africa relative to the Pacific plate is determined by a plate motion chain running from Africa through East Antarctica and West Antarctica to the Pacific (Figs 3, 4).

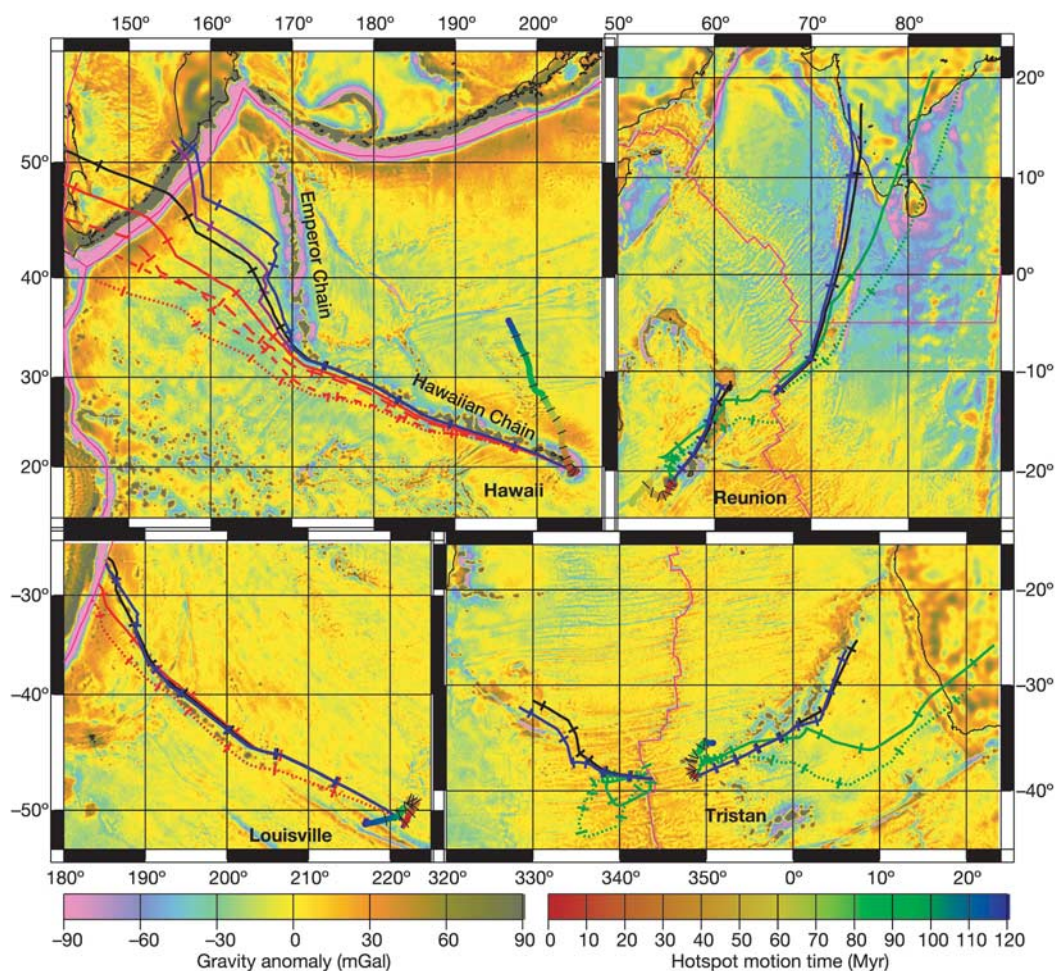


Figure 2 Computed hotspot motion and tracks for the moving-source model. Hotspot motion is shown as rainbow-coloured lines (right colour bar), tracks as single-coloured lines (shown for the past 83 Myr); tickmark interval is 10 Myr for both. Hotspot tracks are computed for the motion of the plate where they are located, except for Hawaii, where tracks computed for motion of the Pacific plate are plotted beyond its boundary. Tracks are plotted regardless of whether a hotspot was actually beneath a given plate at a given time. Red lines (model 1, shown for Hawaii and Louisville), and the purple line (model 2, shown for Hawaii), are computed for absolute plate motions such that the fit to Tristan and Réunion hotspot tracks is optimized for the two plate motion chain models as indicated; optimization parameters are African plate rotations 0–47 and 47–83 Myr ago. Green lines (shown for Réunion and Tristan) are for optimizing the fit to Hawaii and Louisville for plate motion chain model 1; parameters are Pacific plate rotations 0–25, 25–47, 47–62 and 62–83 Myr ago. Black (model 1) and blue (model 2) lines are for optimizing jointly to all

four tracks for the two plate motion chain models as indicated; parameters are African rotations 0–47, 47–62 and 62–83 Myr ago. Dotted lines, hotspots assumed fixed; short-dashed line, only hotspot motion on African hemisphere considered (shown for Hawaii); long-dashed line, only hotspot motion on Pacific hemisphere considered (shown for Hawaii); continuous lines, all hotspot motions considered. A least-squares method⁷ is used to optimize the fit to locations and radiometric ages (see Fig. 5) of seamounts. Free air gravity⁴⁴ shows actual hotspot tracks. This, rather than topography, was chosen because it should better distinguish between actual hotspot tracks (that is, seamounts formed above a plume, in an intraplate location) and material erupted at a ridge through plume–ridge interaction: the former are not expected to be in isostatic equilibrium and should therefore be readily visible on a gravity map. The latter is expected to be approximately in isostatic equilibrium, hence without strong gravity signature.

For times between chrons 24 and 34 (52–83 Myr ago), it is possible to determine Africa–Pacific relative motion by using the same geometrical connection through Antarctica (model 1), but substantial uncertainty surrounds the nature of intra-Antarctic deformation. In model 1 we assume that there was no Antarctic deformation before 43 Myr ago. An alternative plate motion chain through Australia is possible (model 2) and might be more reliable. Australia–East Antarctica motion is known from data in the South-east Indian Ocean²¹. Lord Howe rise–Australia motion is quantified back to an age of chron 33y (73 Myr ago) from Tasman Sea data²². Total Pacific–Lord Howe rise motion since the cessation of spreading in the Tasman Sea (shortly after chron 24) is accurately quantified by fitting rifted oceanic margins south of New Zealand²³. Magnetic anomaly 18 is the oldest anomaly identified in the southeast Tasman Sea²⁴. Stratigraphic evidence, combined with extrapolation of southeast Tasman Sea spreading rates to the rift margin, provides a best fit for the inception of the Eocene Lord Howe rise–Pacific (Campbell plateau) plate boundary at about chron 20^{23,24}. This is also about the time that the Adare basin was formed¹⁹, and the time of most rapid change in the southeast Indian Ocean²¹. For these reasons we choose in model 2 to switch from a West Antarctic relative motion chain at chron 20 to an Australia

relative motion chain at chron 21, and interpolate linearly for times between chrons 20 and 21.

We compute predicted hotspot trails by using both relative plate motion chains and best-fitting ‘absolute’ plate motions in the mantle reference frame⁷. Relative plate motions are listed in

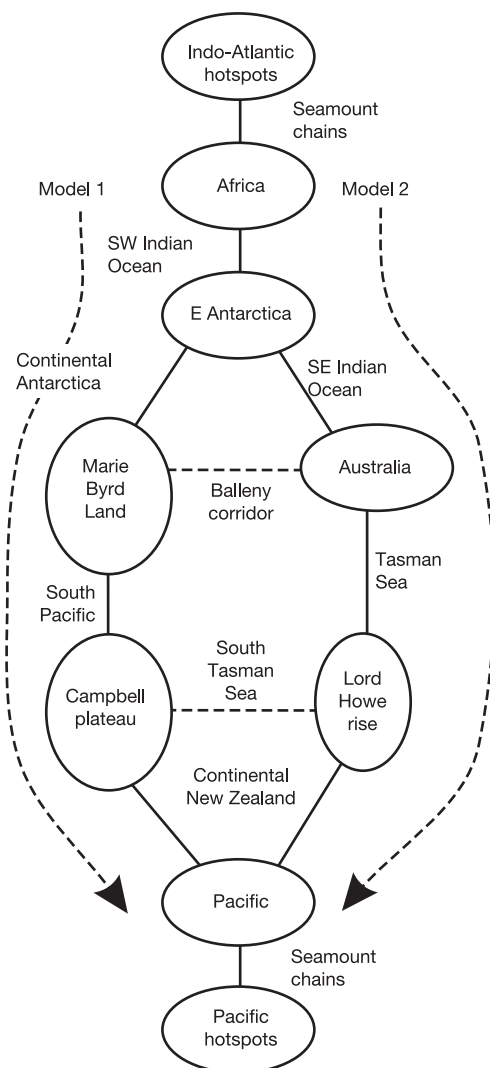


Figure 3 Diagram showing relative plate motion chains considered for times older than chron 20 (43 Myr ago). For times younger than chron 20, models 1 and 2 each follow a chain passing through continental Antarctica. The closed loop in the centre is the South Pacific plate motion circuit. Relative plate motions are listed in Supplementary Table S1.

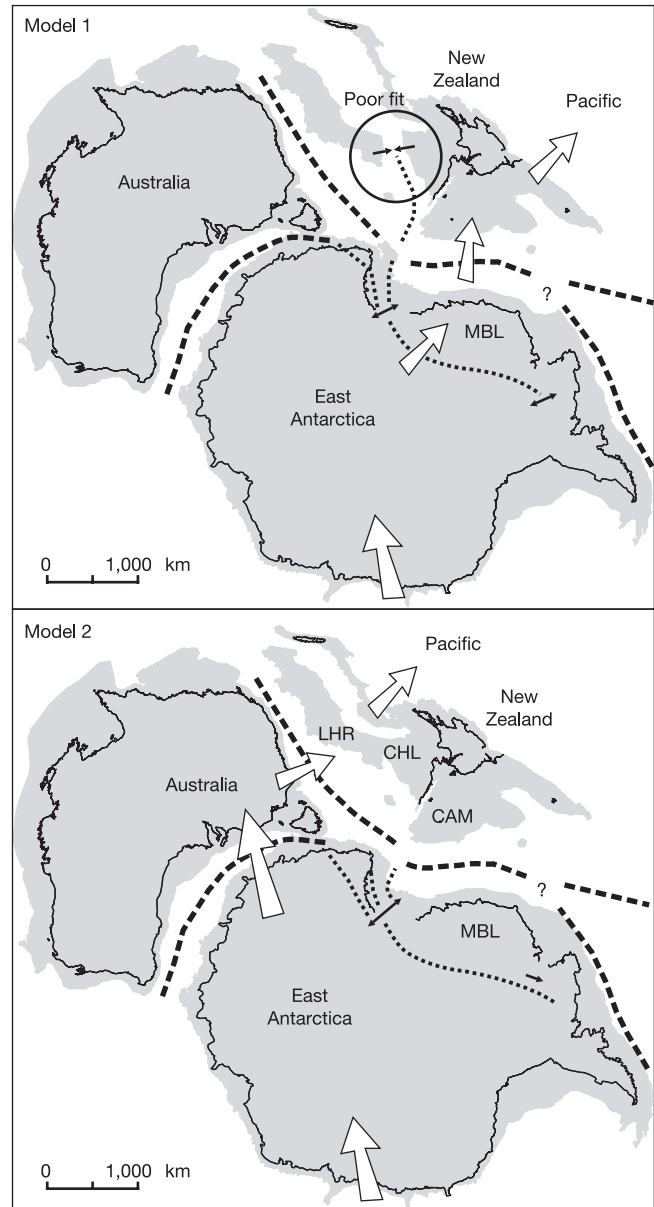


Figure 4 South Pacific reconstructions at chron 31y (68 Myr ago). Present-day coastlines and continental slopes shallower than 2,000 m (shaded) are shown for reference. Open arrows show geographic paths of plate motion chains (Fig. 3). Dashed lines are divergent plate boundaries with seafloor spreading. Dotted lines are inferred intracontinental plate boundaries required to accommodate plate-circuit closure. In model 1, intra-Antarctic motion is specified¹⁹ and residual motion before 43 Myr ago is accommodated within New Zealand. In model 2, residual motion before 43 Myr ago is accommodated within Antarctica. The large convergence required in New Zealand by model 1 during the interval 74–47 Myr ago is inconsistent with geological observations of low sedimentation rate and undeformed stratigraphy overlying rifted basement. In model 2, the total intra-Antarctic plate motion in the Ross Sea is about 300 km of rifting, with required motion becoming smaller and more dextral-oblique to the south and east (finite rotation is 79.7° S, 56.1° W, 6.7°). This is consistent with regional observations. Plate motion arrows in New Zealand and Antarctica are scaled to be double the predicted motion. LHR, Lord Howe rise; CHL, Challenger plateau; CAM, Campbell plateau; MBL, Marie Byrd Land.

Supplementary Table S1. The greatest uncertainty in model 2 lies in the possibility that a plate boundary existed through New Zealand, where there is evidence for minor rifting²⁵. If a through-going Pacific–Lord Howe rise boundary existed and had significant motion, it must have connected to the Pacific–Antarctic ridge southeast of New Zealand. We cannot rule this out, but note that South Pacific seafloor data place a moderately strong constraint on the possible existence of such a boundary for times younger than chron 33y (73 Myr ago) and previous reconstructions of seafloor data have not required one^{18,26}. Deformation in New Zealand is discussed further below.

Model 1: plate motion chain through Antarctica

If we assume fixed hotspots we obtain misfits as previously found^{12,18,27} (dotted red and green lines in Fig. 2). If hotspot motion is considered we obtain the continuous lines. The difference between dotted and continuous lines comes from both hotspot motion in the Pacific (to a greater part) and African (to a lesser part) hemisphere. This is illustrated for the Hawaiian track, for which we also show predictions that separately consider African hotspot motion (short-dashed line) and Pacific hotspot motion (long-dashed line). Hotspot motion in the Pacific and in the African hemisphere tends to move the predicted track in the right direction; for the model shown here, the combined effect is about the right magnitude. Hence we reiterate our previous conclusion¹⁶ that, within the uncertainties of our model, hotspot motion (and in particular, motion of the Hawaiian hotspot) is sufficient to explain any discrepancies between predicted and observed tracks during the

past ~47 Myr, and no ‘missing links’ in the plate motion chain are required.

However, the situation is different before 47 Myr ago, because the predicted Hawaiian track (continuous red line) shows no sign of a bend: predicted locations before 47 Myr ago are too far west. To eliminate that misfit we would need a strong westward component of Hawaiian hotspot motion, in addition to the predicted southward motion. However, none of our models yield such a motion: the computed geometry of large-scale flow in the vicinity of the Hawaiian hotspot, and hence its motion (which is consistently towards the south or southeast in most of our model results) is a consequence of its location between the large upwelling under the Pacific south of Hawaii and regions of past subduction in the north Pacific. Because of the simple geometry of the flow field, there is no reason for a westward component of Hawaiian hotspot motion. The green continuous lines show that computed motion of hotspots in the African hemisphere is equally insufficient to eliminate the misfit: particularly for the Tristan hotspot, a strong additional eastward component of motion would be required before 47 Myr ago, and none of our models yield such a motion—all models yield slow hotspot motion, and not in an eastward direction. The black lines show that a combination of additional westward Hawaiian hotspot motion and eastward Tristan hotspot motion would produce a better fit, but this combination is similarly unlikely.

Model 2: plate motion chain through Australia

Hotspot tracks computed with a plate motion chain through Australia for the interval 83–52 Myr ago are shown as the blue lines in Fig. 2. Although predicted slightly too far west, the northerly trend of the younger part of the Emperor seamount chain is now adequately reproduced, as is the prominent Emperor–Hawaii bend. Palaeolatitudes are predicted well³, but the predicted Hawaiian track before ~65 Myr ago is still west of the actual location of the northern Emperor chain.

For Louisville, Tristan and Réunion, the fit is now very good, and the predicted position of the Tristan track 80 Myr ago is close to the location that presumably marks the ridge crossing²⁸. Tristan and Réunion hotspots have been close to spreading ridges, so that volcanics along their tracks might partly represent flow towards ridges and not eruption above the plume stem^{28,29}. Although uncertainties of hotspot motion cannot be formally quantified, we suggest that any remaining misfit can be attributed to a lateral flow of plume material and inaccuracies associated with the computation of hotspot motions and plate reconstructions. The northern Emperor chain is the only possible exception, where predictions are significantly west of observed locations. Figure 5 shows that this model also yields a good fit to observed age progressions along hotspot tracks, with the exception of the northern Emperor chain. Other misfits might be partly due to inadequate age dates, and might again be partly explained by a lateral flow of plume material and inaccuracies in modelled hotspot motion. For example, a better fit for the Louisville hotspot would be achieved for somewhat faster hotspot motion, still within the range of modelling results obtained¹⁷.

Deformation 65–47 Myr ago

The difference between the predictions of plate motion chain models 1 and 2 is entirely the effect of intraplate deformation not recorded on the ocean floor. For the interval 65–47 Myr ago, the ‘missing’ motion that is required within the closed South Pacific plate-motion circuit is most probably within Antarctica. The implied deformation requires relative motion about a local rotation pole, such that rifting in the Ross Sea diminishes rapidly southwards (Fig. 4). There is abundant evidence for such a deformation.

Geophysical investigations in Antarctica reveal a broad (~1,200-km) region of extended crust adjacent to the Transantarctic Mountains that can be traced through the Ross Sea and southern Marie

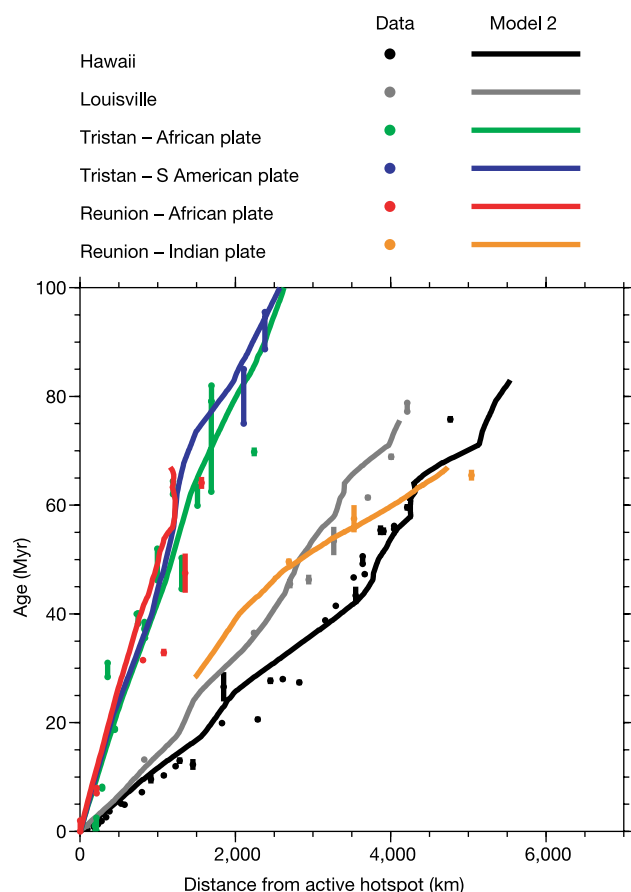


Figure 5 Age progression along hotspot tracks. Lines are for model 2; dots are available age dates, compiled for Hawaii^{3,5,6}, Louisville^{45,46}, Réunion^{47,48} and Tristan^{49,50}. Some of these sources are themselves compilations, which cite original references. Colours represent hotspot tracks as indicated.

Byrd Land^{30,31}. Farther east, the adjacent Thurston Island and Antarctic peninsula regions contain evidence for dextral oblique convergence and subduction-related volcanism that persisted until Oligocene time^{32,33}. The observed geometry of crustal strain is best approximated by a clockwise motion of Marie Byrd Land, relative to East Antarctica, about a local rotation pole near the Weddell Sea³³.

Crustal thickness results from the central Ross Sea imply 350–400 km of extension, which could be revised to 400–450 km if some volcanic addition to the crust were allowed for³⁰. When combined with the Adare trough data, the result suggests that 60–70% of Ross Sea extension occurred before ~50 Myr ago. Seafloor analyses show that some deformation occurred in the northern Ross Sea during the interval 65–47 Myr ago³⁴. Additional motion might have partly transferred to the Tasman–Pacific ridge, through the Emerald Fracture zones. Reconstructions of the Southeast Indian Ocean indicate that deformation might have occurred in the Northern Victorialand–South Tasman rise region²¹. If an alternative hypothesis is considered, namely that deformation was instead within New Zealand, then the required convergent deformation is found to be quite inconsistent with known structure of the region^{25,35–37} (Fig. 4).

Deformation 83–65 Myr ago

Deformation in New Zealand can help to explain the remaining misfit for the oldest part of the Emperor chain: the plate motion chains used in both models 1 and 2 pass through the New Zealand continent and therefore might be affected by unquantified New Zealand deformation. For times younger than chron 27 (61 Myr ago), it has been shown that South Pacific sea floor is consistent with spreading at a single ridge, and hence New Zealand deformation is not significant¹⁸. However, for times older than chron 27, it is known that South Pacific ocean crust is not consistent with a two-plate system¹⁸. In part, this is very likely to be the result of separate motion of the Bellingshausen plate adjacent to Antarctica^{26,38}, which does not affect our conclusions, but some ambiguity exists whether New Zealand deformation is also possible. It has been suggested on geological grounds²⁵. For the interval immediately after the break-up of New Zealand and Antarctica (83–74 Myr ago), there is little doubt that some New Zealand deformation took place^{25,35}. If that deformation is described by a rotation pole anywhere near eastern New Zealand, as suggested by observations of geological and crustal structure, then the effect of including it will be to improve the fit of the older part of model 2. We suggest that at least some of the misfit of the older part of the Emperor seamount chain is associated with as yet unquantified motion within New Zealand, and the subject merits more detailed investigation.

Recent results³⁹ also show that ultraslow spreading is characterized by a lack of transform faults and mantle is emplaced to the sea floor over large regions. Thus, a characteristic magnetic stripe pattern cannot be expected either. It is therefore possible that real uncertainties of plate reconstructions, and real amounts of extension, in regions characterized by very slow divergence are larger than those obtained from formal error analysis. With somewhat larger amounts of Antarctic and New Zealand deformation we could reduce the misfit to an insignificant level. A wide range of models—with deformation axes anywhere within the deformed regions—is adequate to explain hotspot tracks.

The Hawaiian–Emperor bend

Our model 2 explicitly includes a major change in relative plate motions during the interval 43–52 Myr ago. Our computations suggest that, relative to the deeper mantle beneath, the Pacific plate, not the African plate, changed its motion around this time, thus causing the bend in the Hawaiian–Emperor chain¹⁶. The mechanism by which the implied change in force balance was achieved so rapidly remains an important question in geodynamics.

Calculated tractions on the lithosphere, based on a very similar mantle flow model as used for calculating hotspot motion, show a

pattern consistent with the proposed deformation (see Supplementary Information 2). Rapid formation of oceanic lithosphere adjacent to Antarctica after the time of the Emperor–Hawaii bend might be partly responsible for the changing style of Antarctic deformation at around that time. Continental lithosphere deforms more easily than oceanic lithosphere (see, for example, ref. 40 and references therein) and the formation of stronger oceanic lithosphere surrounding Antarctica after ~50 Myr ago might have caused stress concentration and localization of deformation in the Adare basin; this might eventually have caused deformation to stop. The resulting change in the force balance of the Pacific plate, which was smaller at that time, might have been sufficient to initiate subduction in the Izu–Bonin–Mariana trench⁴¹, which might have further accelerated the change in direction of Pacific plate motion.

Discussion

An attempt to explain hotspot tracks globally by intraplate deformation alone, without hotspot motion, is not sufficient, because it would require rotations describing motion within Antarctica or New Zealand that have rotation poles far from the respective regions. Such predictions are not consistent with regional observations. However, the inclusion of hotspot motion associated with the computed plume advection makes it possible to explain hotspot tracks with axes of deformation that lie within the deformed regions, which is consistent with formal uncertainties¹⁹ and geodynamic considerations⁴². Neither hotspot motion nor intraplate deformation alone offers a suitable explanation for the set of global observations; only a combination of the two does so.

The nature of the models and observations makes a formal analysis of error difficult, but the success of the models in reconciling hotspot tracks and plate motions for times since 47 Myr ago, and also the correspondence of the model with palaeolatitudes of older parts of the Hawaiian track³, give it much credence. The case for the cause of the bend in the Hawaiian–Emperor track rests on the comparison of plate motion chains through Australia–Lord Howe rise and East–West Antarctica, where there is a clear discrepancy. This is best explained by East–West Antarctica motion before 47 Myr ago. Such intraplate deformation is roughly consistent with geological evidence and could be associated with additional oceanic deformation that might be hard to identify. Future work might better identify the nature of the deformation in the region and allow testing of the hypothesis presented here. In addition, the models that predict hotspot motion are based on seismic models of mantle heterogeneity, which are limited in resolution, and also on geodynamical assumptions relating seismic properties to density anomalies and mantle flow. These models will certainly improve as better data and more geodynamical constraints, such as reliable images of mantle plumes, become available. Nevertheless, the model we have proposed is plausible, it offers a coherent explanation of observed hotspot tracks, and it is consistent with a wide range of global observations. □

Methods

Figures were prepared using the GMT software⁴³.

Received 4 March; accepted 12 May 2004; doi:10.1038/nature02660.

1. Morgan, W. J. Deep mantle convection plumes and plate motions. *Am. Assoc. Petrol. Geol. Bull.* **56**, 203–213 (1972).
2. Molnar, P. & Stock, J. Relative motions of hotspots in the Pacific, Atlantic and Indian oceans since late Cretaceous time. *Nature* **327**, 587–591 (1987).
3. Tarduno, J. A. *et al.* The Emperor Seamounts: A record of southward motion of the Hawaiian hotspot plume in Earth's mantle. *Science* **301**, 1064–1069 (2003).
4. Besse, J. & Courtillot, V. Apparent and true polar wander and the geometry of the geomagnetic field in the last 200 million years. *J. Geophys. Res.* **107**, 2300 doi:10.1029/2000JB000050 (2002).
5. Clague, D. A. & Dalrymple, G. B. in *The Eastern Pacific Ocean and Hawaii*, vol. N of *The Geology of North America* (eds Winterer, E. L., Hussong, D. M. & Decker, R. W.) 188–217 (GSA, Boulder, CO, 1989).
6. Sharp, W. D. & Clague, D. A. An older, slower Hawaii–Emperor bend. *Eos* **83** (Fall Meet. Suppl.) Abstract T61C-04 (2002).

7. Steinberger, B. Plumes in a convecting mantle: models and observations for individual hotspots. *J. Geophys. Res.* **105**, 11127–11152 (2000).
8. Courtillot, V., Davaille, A., Besse, J. & Stock, J. Three distinct types of hotspots in the Earth's mantle. *Earth Planet. Sci. Lett.* **205**, 295–308 (2003).
9. Steinberger, B. Motion of the Easter Island hotspot relative to hotspots on the Pacific plate. *Geochim. Geophys. Geosyst.* **3**, 8503 doi:10.1029/2002GC000334 (2002).
10. Vink, G. E. A hotspot model for Iceland and the Vøring Plateau. *J. Geophys. Res.* **89**, 9949–9959 (1984).
11. Hofmann, C. *et al.* Timing of the Ethiopian flood basalt event and implications for plume birth and global change. *Nature* **389**, 838–841 (1997).
12. Rochette, P. *et al.* Magnetostratigraphy and timing of the Oligocene Ethiopian traps. *Earth Planet. Sci. Lett.* **164**, 497–510 (1998).
13. Coulie, E. *et al.* Comparative K-Ar and Ar/Ar dating of Ethiopian and Yemenite Oligocene volcanism: Implications for timing and duration of the Ethiopian Traps. *Earth Planet. Sci. Lett.* **206**, 477–492 (2003).
14. Courtillot, V. & Renne, P. On the ages of flood basalt events. *C.R. Geosci.* **335**, 113–140 (2003).
15. Steinberger, B. & O'Connell, R. J. Advection of plumes in mantle flow: implications for hotspot motion, mantle viscosity and plume distribution. *Geophys. J. Int.* **132**, 412–434 (1998).
16. Steinberger, B. & O'Connell, R. J. in *The History and Dynamics of Global Plate Motions*, vol. 121 of *Geophysical Monograph Series* (eds Richards, M. A., Gordon, R. G. & van der Hilst, R. D.) 377–398 (AGU, Washington DC, 2000).
17. Antretter, M., Riisager, P., Hall, S., Zhao, X. & Steinberger, B. in *Origin and Evolution of the Ontong Java Plateau* (eds Fitton, G., Mahoney, J., Wallace, P. & Saunders, A.) Geol. Soc. Lond. Spec. Pub. 229, 21–30 (2004).
18. Cande, S. C., Raymond, C. A., Stock, J. & Haxby, W. F. Geophysics of the Pitman Fracture Zone and Pacific–Antarctic plate motions during the Cenozoic. *Science* **270**, 947–953 (1995).
19. Cande, S. C., Stock, J. M., Mueller, R. D. & Ishihara, T. Cenozoic motion between East and West Antarctica. *Nature* **404**, 145–150 (2000).
20. Nankivell, A. P. *Tectonic Evolution of the Southern Ocean Between Antarctica, South America and Africa over the Past 84 Ma* Thesis, Univ. Oxford (1997).
21. Tikku, A. A. & Cande, S. C. On the fit of Broken Ridge and Kerguelen Plateau. *Earth Planet. Sci. Lett.* **180**, 117–132 (2000).
22. Gaina, C. *et al.* The tectonic history of the Tasman Sea; a puzzle with 13 pieces. *J. Geophys. Res.* **103**, 12413–12433 (1998).
23. Sutherland, R. The Australia–Pacific boundary and Cenozoic plate motions in the SW Pacific; some constraints from Geosat data. *Tectonics* **14**, 819–831 (1995).
24. Wood, R. A., Lamarche, G., Herzer, R. H., Delteil, J. & Davy, B. Paleogene seafloor spreading in the southeast Tasman Sea. *Tectonics* **15**, 966–975 (1996).
25. King, P. R. & Thrasher, G. P. *Cretaceous–Cenozoic Geology and Petroleum Systems of the Taranaki Basin, New Zealand* (Institute of Geological and Nuclear Sciences Monograph 13, Lower Hutt, New Zealand, 1996).
26. Heinemann, J. *et al.* Constraints on the proposed Marie Byrd Land–Bellingshausen plate boundary from seismic reflection data. *J. Geophys. Res.* **104**, 25321–25330 (1999).
27. Raymond, C. A., Stock, J. M. & Cande, S. C. in *The History and Dynamics of Global Plate Motions*, vol. 121 of *Geophysical Monograph Series* (eds Richards, M. A., Gordon, R. G. & van der Hilst, R. D.) 359–376 (AGU, Washington DC, 2000).
28. Sleep, N. H. Ridge-crossing mantle plumes and gaps in tracks. *Geochim. Geophys. Geosyst.* **3**, 8505 doi:10.1029/2001GC000290 (2002).
29. Ebinger, C. & Sleep, N. Cenozoic magmatism throughout east Africa resulting from impact of one large plume. *Nature* **395**, 788–791 (1998).
30. Behrendt, J. C. *et al.* Geophysical studies of the west Antarctic rift arm. *Tectonics* **10**, 1257–1273 (1991).
31. Cooper, A. K., Barker, P. F. & Brancolini, G. E. *Geology and Seismic Stratigraphy of the Antarctic margin*, vol. 68 of *Antarctic Research Series* (AGU, Washington DC, 1995).
32. McCarron, J. J. & Larter, R. D. Late Cretaceous to early Tertiary subduction history of the Antarctic Peninsula. *J. Geol. Soc. Lond.* **155**, 255–268 (1998).
33. Storey, B. C. & Nell, P. A. R. Role of strike-slip faulting in the tectonic evolution of the Antarctic Peninsula. *J. Geol. Soc. Lond.* **145**, 333–337 (1988).
34. Marks, K. M. & Stock, J. M. Early Tertiary gravity field reconstructions of the Southwest Pacific. *Earth Planet. Sci. Lett.* **152**, 267–274 (1997).
35. Sutherland, R. Basement geology and tectonic development of the greater New Zealand region: an interpretation from regional magnetic data. *Tectonophysics* **308**, 341–362 (1999).
36. Laird, M. G. in *South Pacific Sedimentary Basins*, vol. 2 of *Sedimentary Basins of the World* (ed. Ballance, P. F.) 37–49 (Elsevier, Amsterdam, 1993).
37. Cook, R. A., Sutherland, R. & Zhu, H. *Cretaceous–Cenozoic Geology and Petroleum Systems of the Great South Basin, New Zealand* (Institute of Geological and Nuclear Sciences Monograph 20, Lower Hutt, New Zealand, 1999).
38. Stock, J. & Molnar, P. Revised history of early Tertiary plate motion in the south-west Pacific. *Nature* **325**, 495–499 (1987).
39. Dick, H. J. B., Lin, J. & Schouten, H. An ultraslow-spreading class of ocean ridge. *Nature* **426**, 405–412 (2003).
40. Molnar, P. Continental tectonics in the aftermath of plate tectonics. *Nature* **335**, 131–137 (1988).
41. Hall, C. E., Gurnis, M., Sdrolias, M., Lavier, L. L. & Mueller, R. D. Catastrophic initiation of subduction following forced convergence across fracture zones. *Earth Planet. Sci. Lett.* **212**, 15–30 (2003).
42. Zatman, S., Gordon, R. G. & Richards, M. A. Analytic models for the dynamics of diffuse oceanic plate boundaries. *Geophys. J. Int.* **145**, 145–156 (2001).
43. Wessel, P. & Smith, W. H. F. Free software helps map and display data. *Eos* **72**, 441 (1991).
44. Sandwell, D. T. & Smith, W. H. F. Marine gravity anomaly from Geosat and ERS 1 satellite altimetry. *J. Geophys. Res.* **102**, 10039–10054 (1997).
45. Watts, A. B., Weisell, J. K., Duncan, R. A. & Larson, R. L. Origin of the Louisville Ridge and its relationship to the Eltanin Fracture Zone System. *J. Geophys. Res.* **93**, 3051–3077 (1988).
46. Koppers, A. A. P., Duncan, R. A. & Steinberger, B. The implications of a non-linear $^{40}\text{Ar}/^{39}\text{Ar}$ age progression along the Louisville seamount trail for models of fixed and moving hotspots. *Geochim. Geophys. Geosyst.* **5** doi:10.1029/2003GC000671 (2004).
47. Duncan, R. A. & Hargraves, R. B. $^{40}\text{Ar}/^{39}\text{Ar}$ geochronology of basement rocks from the Mascarene Plateau, Chagos Bank and the Maldives Ridge. *Proc. Ocean Drill. Program Sci. Res.* **115**, 43–51 (1990).
48. Hofmann, C., Feraud, G. & Courtillot, V. $^{40}\text{Ar}/^{39}\text{Ar}$ dating of mineral separates and whole rocks from the Western Ghats lava pile: further constraints on the duration and age of the Deccan Traps. *Earth Planet. Sci. Lett.* **180**, 13–27 (2000).
49. O'Connor, J. M. & le Roex, A. P. South Atlantic hot spot–plume systems: 1. Distribution of volcanism in time and space. *Earth Planet. Sci. Lett.* **113**, 343–364 (1992).
50. O'Connor, J. M. & Duncan, R. A. Evolution of the Walvis Ridge – Rio Grande Rise hot-spot system: Implications for African and South American plate motions over plumes. *J. Geophys. Res.* **95**, 17475–17502 (1990).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank P. Molnar, J. Stock, T.H. Torsvik and R.A. Duncan for comments on the manuscript, and N. Sleep for a review. R.S. was supported by the NZ Foundation for Research Science and Technology, the Marsden Fund, and a Caltech Visiting Associate award. R.J.O. received support from the NSF.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.S. (bernhard.steinberger@uni-bayreuth.de).

Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey

Zeev Pancer^{1,2}, Chris T. Amemiya⁶, Götz R. A. Ehrhardt^{1,5}, Jill Ceitlin⁷, G. Larry Gartland^{1,4} & Max D. Cooper^{1,2,3,4,5}

¹Division of Developmental and Clinical Immunology, Departments of ²Medicine, ³Pediatrics and ⁴Microbiology, and the ⁵Howard Hughes Medical Institute, University of Alabama at Birmingham, Birmingham, Alabama 35294, USA

⁶Molecular Genetics Program, Benaroya Research Institute at Virginia Mason, Seattle, Washington, 98101, USA

⁷University of Michigan, Ann Arbor, Michigan 48109, USA

Although jawless vertebrates are apparently capable of adaptive immune responses, they have not been found to possess the recombinatorial antigen receptors shared by all jawed vertebrates. Our search for the phylogenetic roots of adaptive immunity in the lamprey has instead identified a new type of variable lymphocyte receptors (VLRs) composed of highly diverse leucine-rich repeats (LRR) sandwiched between amino- and carboxy-terminal LRRs. An invariant stalk region tethers the VLRs to the cell surface by means of a glycosyl-phosphatidyl-inositol anchor. To generate rearranged VLR genes of the diversity necessary for an anticipatory immune system, the single lamprey VLR locus contains a large bank of diverse LRR cassettes, available for insertion into an incomplete germline VLR gene. Individual lymphocytes express a uniquely rearranged VLR gene in monoallelic fashion. Different evolutionary strategies were thus used to generate highly diverse lymphocyte receptors through rearrangement of LRR modules in agnathans (jawless fish) and of immunoglobulin gene segments in gnathostomes (jawed vertebrates).

Adaptive immune responses in jawed vertebrates are initiated when antigens are recognized by specific lymphocyte receptors. Antigen receptor diversity is generated through recombination of variable, diversity and joining gene segments in the immunoglobulin and T-cell receptor (TCR) gene loci. This combinatorial rearrangement generates vast repertoires of antibodies against unprocessed antigens and of TCRs that recognize antigen fragments presented by major histocompatibility complex (MHC) class I and II molecules. Clonally diverse lymphocytes thus form the cornerstone of vertebrate adaptive immunity in the form of immunoglobulin-bearing B cells and TCR-bearing T cells that differentiate from stem-cell precursors within primary haematopoietic tissues and the thymus. The principal elements of this recombinatorial immune system are conserved in all jawed vertebrates, and the multigene TCR and immunoglobulin loci are remarkably complex even in the most basal gnathostome representatives, such as sharks, skates and rays^{1–3}.

Adaptive immune responses have also been reported in the jawless vertebrates lamprey and hagfish, the only surviving descendants from the early vertebrate radiation⁴. For example, lampreys produce specific circulating agglutinins in response to immunization^{5–10}, reject second set skin allografts at an accelerated rate^{9,11} and exhibit delayed type hypersensitivity reactions⁵. These responses have been attributed to agnathan cells that morphologically resemble the lymphocytes found in jawed vertebrates^{5,9,11–16}. As with their mammalian counterparts, lamprey lymphocytes are more irradiation-sensitive than other blood cell types⁹, aggregate and proliferate in response to antigenic stimulation^{5,12}, and express transcription factors such as PU.1/Spi-B and Ikaros that are involved in mammalian lymphocyte differentiation^{17–20}. However, immunoglobulin, TCR and MHC genes have not been identified in jawless vertebrates nor in the draft genome sequence of the invertebrate urochordate *Ciona intestinalis*²¹.

With this in mind, we initiated a search for primordial molecular elements of the vertebrate immune system in the sea lamprey *Petromyzon marinus*. In an earlier analysis of transcripts expressed by lymphocyte-like cells from lamprey haematopoietic tissues we identified several homologues of immune system molecules, but

none of the immunoglobulin superfamily receptor elements used by jawed vertebrates for specific adaptive immunity^{16,22,23}. Reasoning that immunocompetent cells in the circulation would be more likely to express the genes involved in adaptive responses, we began with a survey of the transcriptome of activated lymphocytes from the blood of immunostimulated lamprey larvae. This search revealed a new type of highly variable lymphocyte receptors.

Immunostimulated blood lymphocyte transcripts

To survey the transcriptome of activated lymphocytes, lamprey larvae were stimulated by intraperitoneal injections of an antigen/mitogen cocktail two–four times at weekly intervals. The fraction of large lymphocytes among peripheral blood leukocytes 3 days after the second booster stimulation was 13-fold greater than in unstimulated individuals, and the fraction of myeloid cells was sixfold greater (Fig. 1a). Compared with the small lymphocyte-like cells, the activated lymphocytes were nearly double in size, had extensive azurophilic cytoplasm and featured prominent nucleoli (Fig. 1b). These cells were sorted on the basis of their light scatter characteristics and used to construct subtracted complementary DNA libraries enriched in messages of activated lymphocytes.

The most abundant group of sequences identified among 1,507 subtracted clones consisted of 319 proteins featuring variable numbers of diverse LRR motifs, which clustered with a set of 52 LRR-containing expressed sequence tags (ESTs) from our database of unstimulated lymphocyte-like transcripts. A set of 239 uniquely diverse LRR proteins was obtained after purging sequences with only the 3' ends, and 22 ESTs from this set encoded most or all of the open reading frames (ORFs) of 239–304 residues in length (Supplementary Fig. 1). These proteins were provisionally named variable lymphocyte receptors (VLR) because each of these 239 sequences was unique and their transcripts were found to be expressed predominantly or exclusively by lymphocytes (Fig. 1c). Unstimulated animals showed highest VLR levels in lymphocytes from haematopoietic tissues, whereas immune stimulation resulted in enhanced VLR transcription by the large blood lymphocytes. The VLR transcripts observed in the myeloid lane of Fig. 1c may reflect

an approximately 5% contamination with activated lymphocytes.

The basic composition of VLRs includes a conserved signal peptide, N-terminal LRR (LRRNT), a variable number of diverse LRRs, a connecting peptide followed by a C-terminal LRR (LRRCT), and a conserved C terminus composed of a threonine/proline-rich stalk, a glycosyl-phosphatidyl-inositol (GPI)-anchor site and a hydrophobic tail (Fig. 1d; see also Supplementary Fig. 2). When a recombinant epitope-tagged VLR was expressed in a mammalian cell line, immunofluorescence analysis confirmed the cell-surface localization of the protein. Treatment with bacterial GPI-specific phospholipase C significantly reduced the level of surface VLR (Fig. 1e) and released the protein into the supernatant. The diversity region from the longest VLR sequence, consisting of 12 LRR modules, was threaded on the crystal structure coordinates of structurally related LRR proteins to generate a three-dimensional model²⁴. This model predicts a concave solenoid structure capped on both ends by the LRRNT and LRRCT, in which LRRs 1–9 form β -sheets and the connecting peptide forms an α -helix (Fig. 1f), resembling the model predicted for Toll-like receptor (TLR) ectodomains²⁵.

Individually diverse VLR repertoires

Blood leukocyte messenger RNA from three immunostimulated and four unstimulated larvae was used for polymerase chain reaction with reverse transcription (RT-PCR) amplification of

VLRs with primers flanking the diversity region (Supplementary Table 1). Sequencing 9–11 clones per animal yielded 69 unique VLRs and only one case of two identical clones from one individual. VLR sequences from two animals are shown to illustrate the protein diversity (Fig. 2; entire set included in Fig. 3 and Supplementary Fig. 3). The size variation of 129–202 residues is primarily due to differences in the number of LRR modules. Each sequence contains an LRRNT of 30–38 residues, an 18-residue LRR1, 1–9 LRRs almost invariably consisting of 24 residues, a 13-residue connecting peptide and a C-terminal LRRCT of 48–58 residues. Regions of pronounced sequence diversity are evident for each LRR motif, almost always consisting of non-similar residues, yet the first seven residues in the LRRNT and the last 20 residues in the LRRCT are nearly invariant.

To assess VLR diversity at the level of individual lymphocytes we used RT-PCR with two sets of nested primers flanking the entire ORF. Analysis of the amplicons obtained from six single lymphocytes sorted from the blood of one unstimulated animal and seven lymphocytes from an immunostimulated larva showed that 12 of the 13 lymphocytes expressed a single VLR (Fig. 3; animals 8, 9). Only one cell isolate yielded two VLR sequences (9.16S, 9.16L), and this isolate might have contained two lymphocytes. Control reactions consisted of four cDNA mixture pairs from the first round of eight single-lymphocyte RT-PCR reactions; subsequent amplification with the nested primers yielded two amplicons for each reaction. In addition, five out of six VLR clones from a pool of

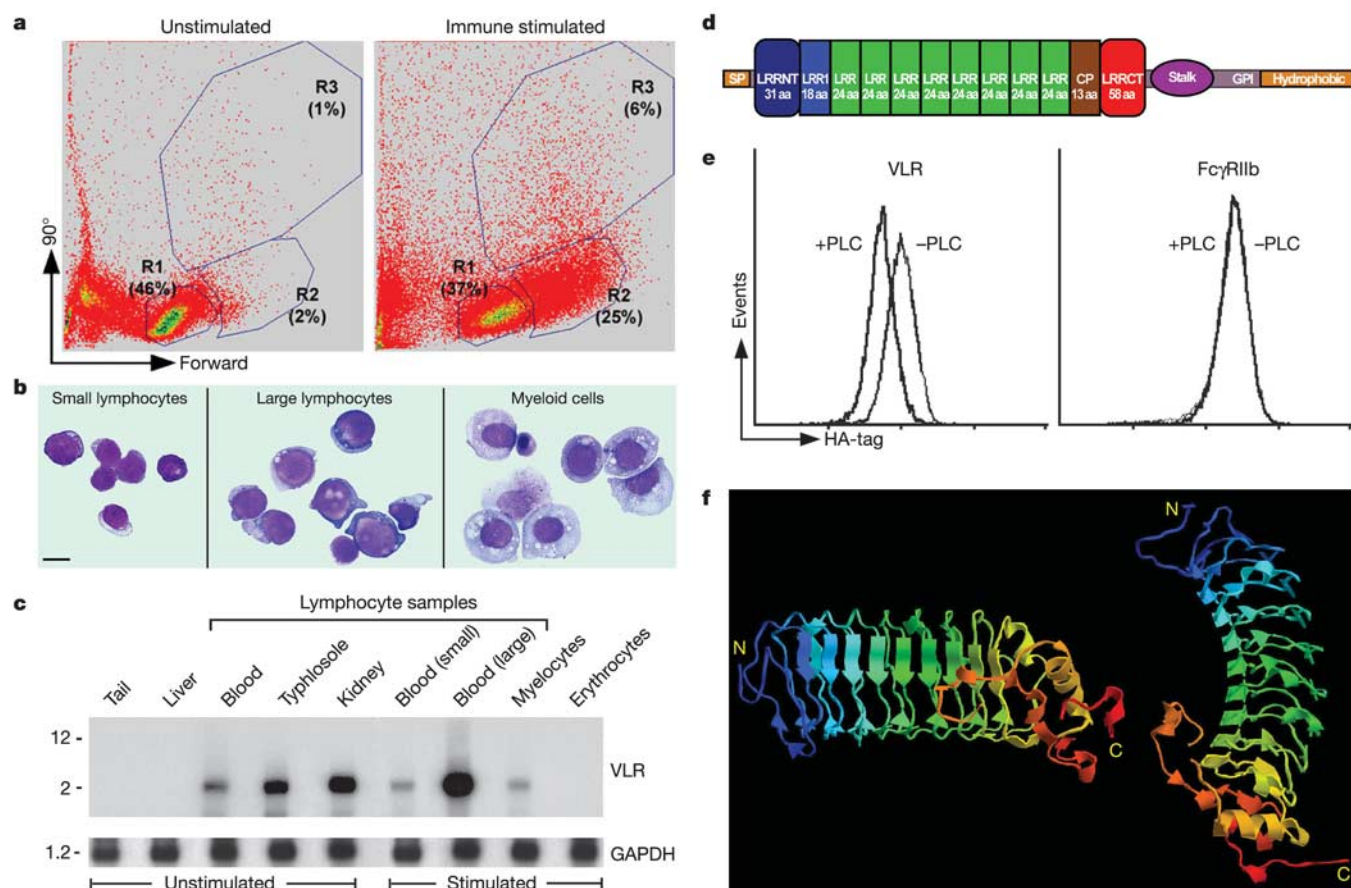


Figure 1 Lamprey leukocytes and VLRs. **a**, FACS analysis of blood leukocytes before and after antigen/mitogen cocktail immunostimulation. **b**, Immunostimulated leukocytes. Small lymphocytes, large lymphocytes and myeloid cells correspond with R1, R2 and R3 respectively, in **a**. Wright-Giemsa stain. Scale bar, 10 μ m. **c**, VLR Virtual northern blot. Amplified cDNA from larval tissues, or sorted cells from unstimulated and immunostimulated blood and haematopoietic organs. Molecular size is indicated in

kilobases. **d**, VLR stick model. Sections from left to right are: signal peptide, N-terminal LRR, nine LRRs, connecting peptide, C-terminal LRR, threonine/proline-rich stalk, GPI anchor and hydrophobic tail (clone 12.26, AY577974). **e**, Epitope-tagged VLR and Fc γ RIIb (control) expressed in mouse thymoma cells, treated with (+PLC) or without (–PLC) GPI-phospholipase C. **f**, Three-dimensional model of VLR diversity region (two rotations).

ten unstimulated cells were unique. These controls indicate that if more than one VLR template were present in a single lymphocyte sample, these would be amplified simultaneously. Notably, combinations of identical VLRs were identified among five lymphocytes from the immunostimulated larva (VLR 9.1 = 9.16S; 9.2 = 9.16L; 9.7 = 9.9; See Fig. 3). These findings are indicative of monoallelic expression of the VLRs, and provide preliminary evidence for clonal expansion of VLR-bearing lymphocytes.

Complexity of the VLR locus

Genome blot hybridization with a C-terminal VLR probe revealed a single band for all three digests used (Fig. 4a). The N-terminal probe, consisting of the 5' untranslated region (UTR) and signal peptide (Fig. 5a), hybridized with 2–3 bands depending on the restriction enzyme used, except for one individual whose blot showed two additional *Bam*HI bands. For high-resolution genome spread we probed a pulse-field blot of erythrocyte DNA (Fig. 4b). In all six digests a single hybridizing band was detected with the C-terminal probe, and the N-terminal probe produced a matching pattern with one additional 350-kilobase *Not*I band. These findings indicate a single VLR locus, with the N terminus and C terminus of the VLR gene (germline VLR; *gVLR*) being contained within 100–150 kb of the genome (Fig. 4b; *Pac*I digest). To characterize further the locus, these probes were used to screen a large insert sea lamprey P1 bacterial artificial chromosome (PAC) library constructed from erythrocyte DNA of one adult. In an analysis of the five PACs that hybridized with both probes, a single 14-kb *gVLR* amplicon was identified by long-range PCR (LR-PCR). Restriction-enzyme analysis of these amplicons revealed identical *Eco*RI bands and two allelic *Bam*HI patterns. PAC clones representing the two *gVLR* alleles were sequenced (PAC3 and PAC16) with inserts of 33 and 44 kb, respectively (Fig. 5a). Their sequences overlapped a 20-kb region containing *gVLR*, and furthermore, PAC16 extended 25 kb upstream from *gVLR* whereas PAC3 extended 18 kb downstream. The overlap region between PAC3 and PAC16 was identical, except for short deletions in *gVLR* from PAC16 (24, 43 and 78 bp). These

sequences were therefore melded into a *gVLR* contig preserving the slightly longer sequence of PAC3.

gVLR in the PAC3/16 contig consists of four exons. The first contains part of the 5' UTR; exon 2 contains the rest of the 5' UTR, a signal peptide and the 5' half of LRRNT; exon 3 encodes the 5' half of LRRCT; and exon 4 encodes the 3' half of LRRCT, the C terminus and 3' UTR. Canonical eukaryotic splice sites were identified only in the 5' UTR intron, whereas other exon–intron boundaries in *gVLR* were determined by alignment to cDNA sequences. Notably, the *gVLR* sequence did not contain a 3' LRRNT, LRR1 or any of the 24-residue LRRs. Upstream from *gVLR*, six cassettes of singlet or doublet variable LRR modules were identified, including LRRNTs, LRR1 and 24-residue LRRs positioned either in forward or reverse orientation. These LRR cassettes spanned the first 6 kb of the contig, whereas two diverse 5' LRRCT cassettes were located 7 kb downstream from *gVLR*.

A survey of PAC clones that hybridized only with the N-terminal probe—by PCR with LRRNT and LRR1 consensus primers—indicated that PAC4 encoded multiple LRRs. The entire insert was 58 kb long (Fig. 5b) and its sequence aligned to 11.7 kb of the *gVLR* contig with minor gaps (four gaps of 210–738 bp in PAC4; eight gaps of 25–55 bp in the PAC3/16 contig). The overlap extended into the intervening sequence between *gVLR* exons 2 and 3, but the 553-bp terminal sequence of PAC4 was unique. Thirty diverse LRR modules, arranged in 17 cassettes of 1–3 modules, were identified in a 31-kb region located 15-kb upstream from the partial 5' *gVLR*. Sequence comparison between PAC4 and the PAC3/16 contig revealed additional 1–5-kb regions with >90% identity, but these were disrupted by unrelated sequences. Hence, the 3' terminal ~12 kb of PAC4 might represent either a duplication of about half of the *gVLR* and the 5' flank, or a highly divergent VLR allele. To estimate the level of polymorphism in the VLR locus we compared the pattern of hybridization in the N-terminal Southern blot to the map of restriction sites in the *gVLR* genes from all three PAC inserts. The pattern of bands and restriction maps was compatible except for a 5.7-kb PAC4 *Hind*III fragment instead of the 2-kb band in the

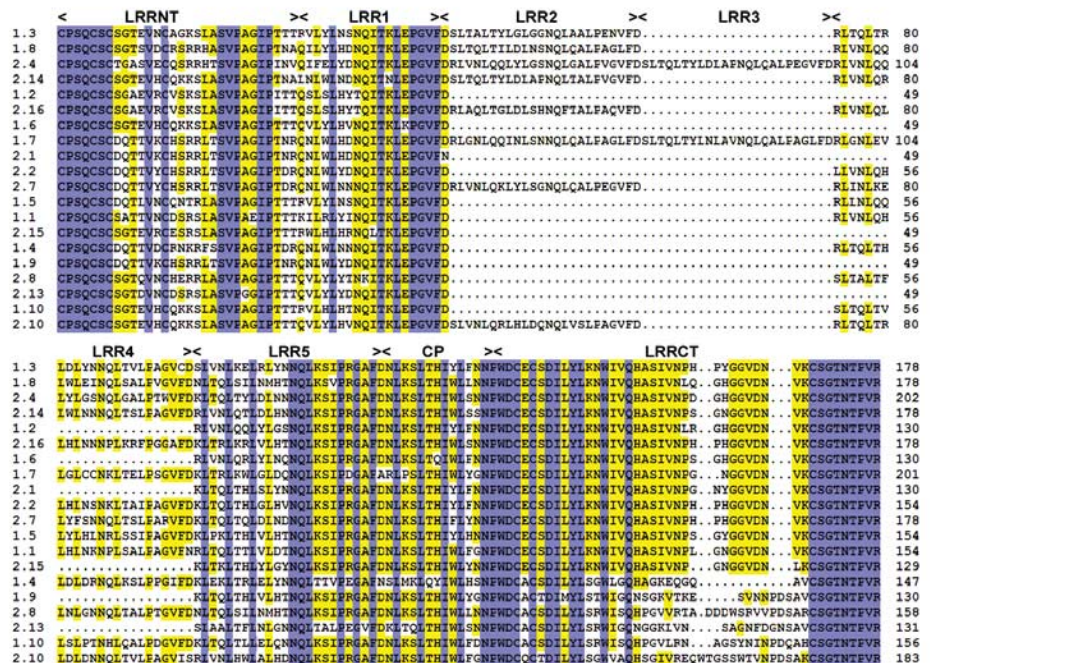


Figure 2 Diversity of VLRs from two lampreys. Alignment of 20 diversity regions, from LRRNT to near C terminus of LRRCT, PCR amplified from blood leukocyte mRNA. Locations of LRR motifs, clone number (for example, 1.3 indicates animal 1, clone 3) and

length are indicated. Blue, 100% identity; yellow, 60–99%; white, <60%. All substitutions are non-similar residues.

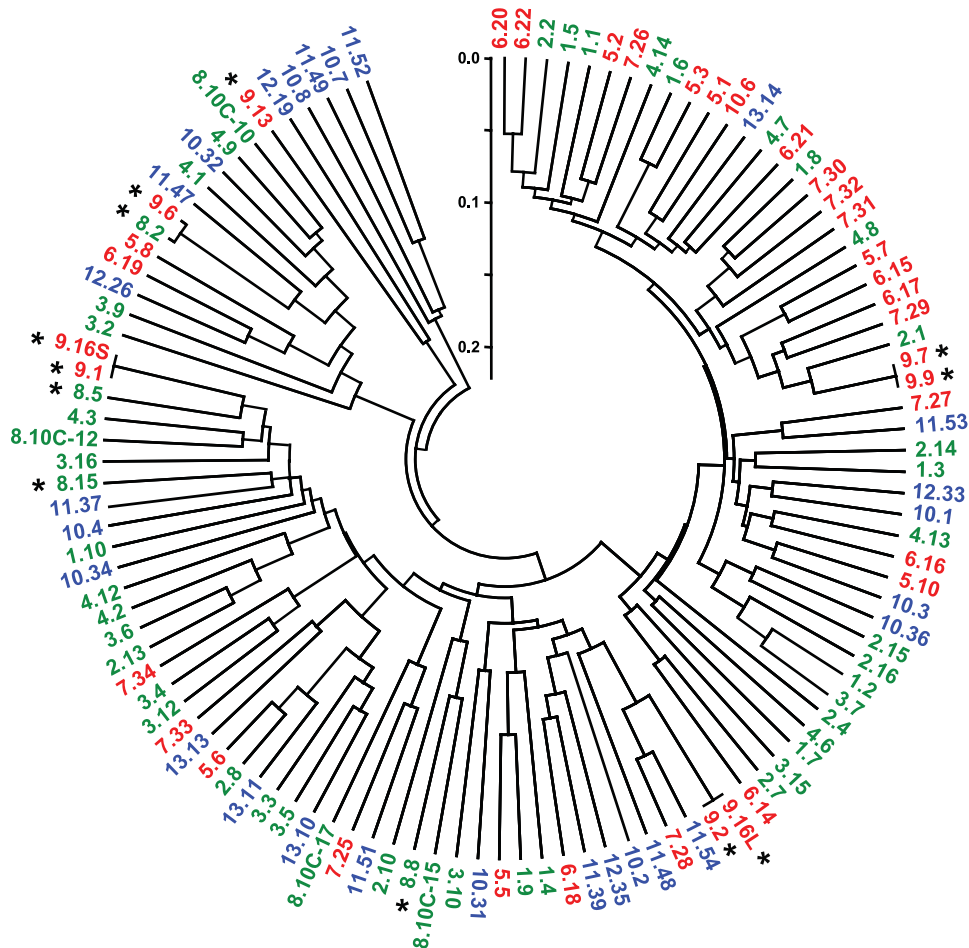


Figure 3 Graphic representation of protein divergence among 112 VLR diversity regions PCR amplified from 13 individual lampreys. Animal and clone numbers are indicated in red for immunostimulated ($N = 27$) and green for unstimulated ($N = 41$) larvae. Asterisks indicate VLRs derived from single lymphocytes ($N = 12$), including two VLRs from one isolate (9.16S, 9.16L); clones derived from a control ten-cell pool are denoted

10C ($N = 4$). Mature VLR sequences derived from genomic DNA are in blue ($N = 28$; blood, animal 10, 12; carcass, animal 11, 13). The mean diversity is 0.36 ± 0.03 ($\pm$ s.d.), ranging from 0.28 to 0.54 within the groups of sequences from 13 individuals. Scale bar indicates per cent amino acid divergence.

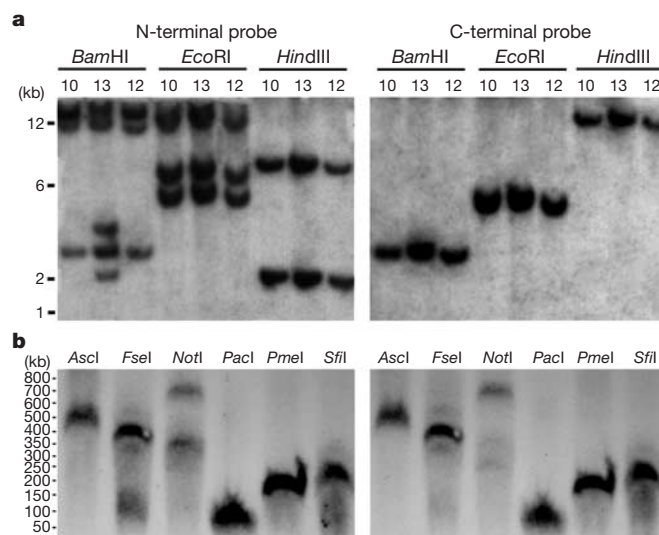


Figure 4 Sea lamprey genome harbours a single VLR locus. Genome blots of restriction-enzyme-digested DNA were hybridized with VLR N-terminal or C-terminal probes. **a**, Blots of three lampreys (blood DNA, animal 10, 12; carcass, animal 13). Only animal 13 shows

a polymorphic *Bam*HI pattern. **b**, Genome spread of erythrocytes pooled from ten lampreys. Pulse-field blot hybridization shows matching patterns for both probes, with an additional 350-kb *Not*I N-terminal band corresponding to a 5' *gVLR* duplication.

Southern blot (Fig. 4a). As only limited polymorphism was observed in an analysis of *gVLR* from three larvae and the PAC library donor, it is unlikely that the 5' half of *gVLR* from PAC4 represents a divergent allele. The limited genetic diversity indicated here for the *VLR* locus is consistent with that reported for micro-satellite loci²⁶, thus indicating a high level of inbreeding among the recently landlocked population of sea lamprey in the North American Great Lakes. We conclude from this analysis that a 5' half *gVLR* duplication has occurred within ~100 kb of *gVLR*.

Somatic rearrangement generates mature VLRs

PCR amplification from blood or larval carcass DNA with primers flanking the *VLR* diversity region produced intronless *VLR* ORFs of 0.5–0.8 kb. PCR with two other sets of ORF-flanking primers produced similar amplicons of ~1 kb, or ~1.7 kb including the 5' UTR intron. Analysis of these PCR products revealed unique *VLR* sequences in all 19 clones (Fig. 3; numbers 10–13). Because these genomic amplicons encoded uninterrupted *VLR* ORFs, which contained all the diversity LRR modules identified in *VLR* transcripts, they were provisionally named mature VLRs to distinguish them from the 'incomplete' *gVLR*. Sequence analysis indicated that the ~1.7-kb mature VLRs should be represented by polymorphic *EcoRI* bands of 1–1.5 kb in the N-terminal Southern blot, but none of these genomic DNA samples yielded visible *EcoRI* bands of <5 kb (Fig. 4a). These observations indicate that mature *VLR* templates are included in DNA samples extracted from larval carcasses or blood, yet paradoxically only *gVLR* was detected in DNA blots from these samples.

To address this issue we proposed that somatic gene rearrangement¹ in lamprey lymphocytes might generate the mature VLRs by insertion of diverse LRR modules from upstream and downstream cassettes into *gVLR*. Primers were therefore designed for PCR amplification across *gVLR*, including flanking regions ~3 kb upstream and ~3 kb of *gVLR* (Fig. 5a). LR-PCR amplification from genomic DNA samples yielded a minor band of ~20 kb, similar to the *gVLR* amplicon from PAC16, plus a prominent band of ~8 kb (Fig. 5c). Sequence analysis of the 8-kb amplicons from two larval DNA samples revealed unique mature VLRs in nine out of

ten clones (Fig. 3; animal 10, blood; animal 13, carcass—two identical clones), the flanks of which were identical to those of *gVLR* (Fig. 5e). In contrast, LR-PCR amplification with primers flanking the *gVLR* ORF from 100-cell pools of erythrocytes or lymphocytes revealed the 14-kb *gVLR* amplicon in both, whereas the ~1-kb mature VLRs were observed only in lymphocytes (Fig. 5d). We conclude that mature *VLR* amplicons from blood or carcass DNA are derived from the lymphocytes in these samples, with preferential amplification favouring the shorter mature *VLR* templates over the more abundant but much longer *gVLR* templates. Altogether 28 unique mature VLRs were identified among random samples from four lamprey genomic amplicons, thus indicating a level of diversity equivalent to that found in *VLR* transcripts.

Discussion

Our search for receptors that could mediate adaptive immune responses in the lamprey has identified a novel system of variable lymphocyte receptors. The VLRs consist of multiple LRR modules and an invariant stalk region that is attached to the lymphocyte plasma membrane by a GPI anchor. The flanking ends of the N-terminal and C-terminal LRRs are invariant and the remarkable *VLR* diversity derives from the variation in sequence and number of the LRR modules. Our initial assessment identified unique transcripts in 345 out of 354 sequences, three pairs of identical VLRs from immunostimulated lymphocytes and only three nearly identical sequences. Random samples of *VLR* clones from individual lampreys also revealed a prevalence of unique sequences; for example, all 12 *VLR* clones from animal 10 were unique. This agnathan representative is thus endowed with a vast repertoire of lymphocyte receptors that is expressed predominantly in a mono-allelic fashion.

A single germline *VLR* gene (*gVLR*) was identified in the lamprey genome. It comprises four exons that encode only the signal peptide, 5' LRRNT, 5' LRRCT, 3' LRRCT and the C terminus. Lacking in diversity LRR modules, this incomplete *gVLR* could not encode the highly diverse *VLR* transcripts. However, clusters of diverse LRR cassettes were found upstream and downstream from

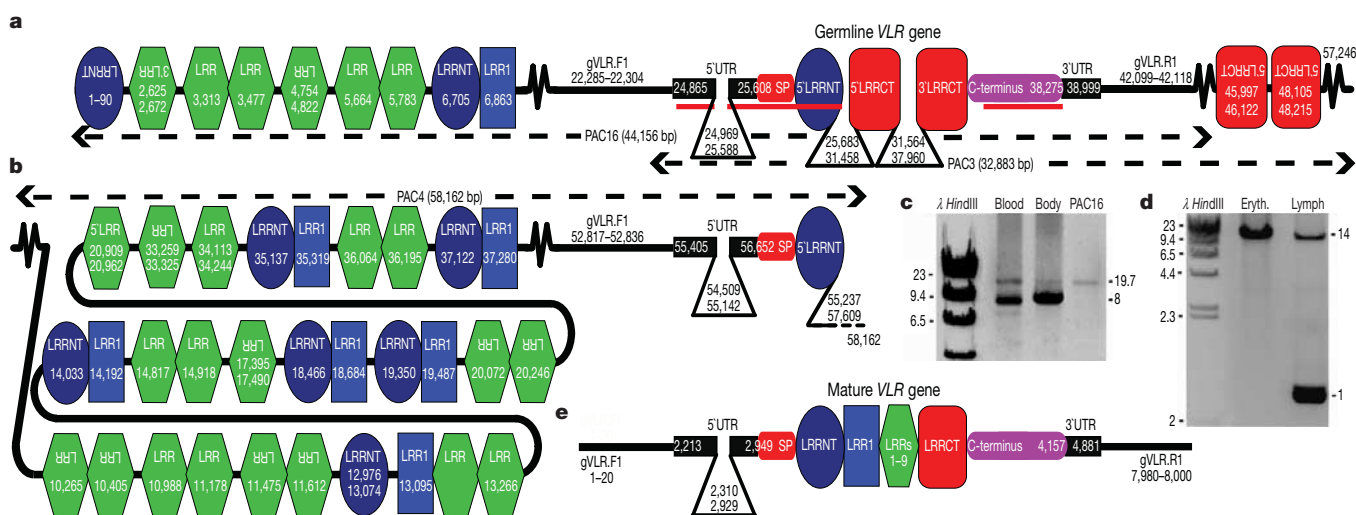


Figure 5 The *VLR* locus. **a**, Motifs in the 57-kb *gVLR* contig (PAC16, 44 kb; PAC3, 33 kb; AY577941). Cassettes of 1–2 LRRs are positioned in forward or reverse orientations. Dashed lines represent PAC inserts; red bars indicate N-terminal and C-terminal probes; gVLR.F1 and gVLR.R1 indicate LR-PCR primers. **b**, Motifs in PAC4 (58 kb, nucleotides 45,882–57,609, align with the *gVLR* contig; AY577942). **c**, LR-PCR analysis of *gVLR*, amplified from blood DNA (animal 10), carcass (animal 13) or PAC16 (control). The

germline *VLR* is ~20 kb whereas the mature VLRs are ~8 kb. **d**, Lymphocyte-specific rearrangement of mature VLRs. LR-PCR with ORF-flanking primers from pools of 100 lymphocytes or erythrocytes: germline *VLR* ~14 kb; lymphocyte mature VLRs ~1 kb. Numbers along left and right sides of panels in **c** and **d** indicate molecular size in kilobases. **e**, An 8-kb mature *VLR* amplicon.

gVLR, and these would be available for insertion into *gVLR* to assemble mature VLR genes. To test the hypothesis that mature VLRs are generated through insertion of upstream and downstream LRR cassettes, which replace non-coding DNA in the lymphocyte *gVLR*, we used LR-PCR to detect both germline and mature VLR genes in lamprey genomic DNA from blood and carcasses. Two amplicons resulted: the expected *gVLR* product of ~20 kb and the predicated ~8-kb amplicon that consisted of a diverse set of mature VLRs. In contrast, when purified populations of erythrocytes and lymphocytes were examined, mature VLR amplicons were found only in the lymphocytes. The mature VLRs are thus generated through a process of somatic DNA rearrangement in lymphocytes.

The highly diverse VLRs may serve a role in recognition of pathogens. Proteins featuring heterogeneous LRR modules are principal innate immune receptors of animals and plants owing to their ability to interact with a wide range of pathogen-associated ligands. For instance, the LRR ectodomains of vertebrate TLRs are implicated in recognition of conserved epitopes of viruses, bacteria, fungi and protozoa, thereby activating signal transduction cascades that culminate in inflammatory responses²⁷. CD14, a GPI-anchored LRR protein that is also found in a soluble form, binds bacterial lipopolysaccharide and phospholipids to form a signalling complex with the TLR4 receptor²⁸. Yet another mammalian family of cytosolic LRR proteins, the NBS-LRRs, recognize intracellular pathogens²⁹. Plant disease resistance genes are members of large multigene families including hundreds of NBS-LRR proteins, LRR-receptor-like kinases and LRR-receptor-like proteins, many of which are involved in specific activation of anti-pathogen responses³⁰. In the likely event that VLRs prove to be capable of binding antigens, their remarkable diversity makes them excellent candidates to mediate the humoral immune responses observed in lampreys. The GPI anchorage of VLRs to the surface of lymphocytes might allow GPI-specific phospholipase release of these receptors³¹, thereby endowing VLRs with dual functionality, both as cell-surface receptors and humoral agglutinins in an anticipatory immune system.

Many features of the lamprey VLR system bear analogy to the variable B and T lymphocyte receptors of jawed vertebrates, with the notable difference that lamprey VLRs consist of LRR modules whereas gnathostome antigen receptors consist of immunoglobulin domains. Lamprey immunity may therefore represent a traceable evolutionary process in which the ancestral strategy of multigene families of heterogeneous LRR receptors gave rise to a novel system of variable lymphocyte LRR receptors that are somatically diversified versions of the single germline VLR gene. In contrast, the deployment of immunoglobulin domains as core components of jawed vertebrate recombinatorial lymphocyte receptors represents an intriguing although as yet untraceable evolutionary innovation, as immune recognition of pathogens and allografts by means of immunoglobulin superfamily members have been shown only in the jawed vertebrates³².

Elucidation of the precise mechanism underlying the somatic generation of mature VLRs will require further study, but we speculate that the conserved flanking ends in the *gVLR5'* LRRNT and 3' LRRCT might serve as anchoring regions for a gene conversion process similar to that elucidated for avian antibody diversification^{33,34}. Indeed, candidate LRR-module donors could be identified among the *gVLR* neighbouring cassettes on the basis of partial identity to VLR sequences. Furthermore, rearrangement of genes encoding surface components is a strategy often used by pathogens to evade immune recognition during chronic infection. Antigenic variation in the pilin of *Neisseria gonorrhoeae* involves non-reciprocal recombination between multiple genes³⁵, and surface antigen variation in *Borrelia spirochaetes*, the causative agent of Lyme disease, is generated by gene conversion between an array of 15 silent cassettes and one expression site³⁶. Recurring DNA rearrangements are also used by the protozoan *Trypanosoma brucei*

for alternative expression of its surface coat glycoprotein genes³⁷, and a similar mechanism may drive the frequent switching among multiple surface antigens in the malaria parasite *Plasmodium falciparum*³⁸ and the intestinal dweller *Giardia lamblia*³⁹. In the evolutionary arms race between hosts and parasites, vertebrates have adopted a similar strategy to combat infectious disease by somatic rearrangement of germline receptor units. The diverse lymphocyte antigen receptors are assembled by means of the cut-and-paste activity of the paired transposase-like RAG1/RAG2 in gnathostomes⁴⁰, and via a mechanism that we are just beginning to unravel in Agnatha. □

Methods

Animals, immune stimulation and blood leukocytes

Sea lamprey larvae, 8–13 cm long, from tributaries to Lake Michigan (Lamprey Services) or Lake Huron (Hammond Bay Biological Station), were sedated (100 mg l⁻¹ MS222; Sigma) for intraperitoneal injection with 75 µl 0.67 × PBS containing: 10⁷ live *Escherichia coli* bacteria, 10⁷ sheep erythrocytes, 50 µg phytohaemagglutinin and 25 µg pokeweed mitogen (Sigma). Larvae were injected two or four times at weekly intervals and cells were collected 3–4 days later. Blood was drained from tail-severed larvae and diluted 1:1 with 0.57 × PBS, 30 mM EDTA. Buffy coat leukocytes were collected after 5 min centrifugation at 50 g and sorted using a MoFlo cytometer (Cytomation).

Subtracted lymphocyte cDNA libraries

The Super SMART PCR cDNA Synthesis kit (BD Biosciences) was used with mRNA from activated blood lymphocytes, myeloid cells and erythrocytes sorted from 21 larvae immunostimulated four times. Activated lymphocyte cDNA was subtracted in two reactions, against myeloid or erythrocyte cDNA (PCR-Select, BD Biosciences). Subtracted products were cloned in pGEM-T (Promega) and 1,507 sequences were analysed.

VLR RT-PCR

All PCR primers are listed in Supplementary Table 1. Buffy coat leukocytes from unstimulated larvae (numbers 1–4), or larvae that had been immunostimulated twice (numbers 5–7), were pelleted for 5 min at 300 g. VLR diversity regions were amplified from random primed cDNA (SuperScript III; Invitrogen) with Expand High Fidelity (Roche) using LRR.F1 + LRR.R1. Cycling parameters were: 94 °C for 1 min, then 35 cycles of 94 °C for 30 s, 59 °C for 30 s, 72 °C for 1 min.

Single lymphocytes (*N* = 32), or ten-cell pools of buffy coat leukocytes from an unstimulated larva (animal 8) and one twice-immunostimulated larva (animal 9), were sorted into 0.2 ml TRIzol (Invitrogen) and cDNA was primed with LRR_C.R2. VLRs were amplified by two rounds of nested PCR—first Slit.F + LRR_C.R2 (Advantage II; BD Biosciences) then LRR_N.F1 + LRR_C.R1—using Expand High Fidelity. Colony PCR with vector primers revealed single-sized inserts in six colonies from each of 12 lymphocytes, three of which were sequenced. Colonies from isolate 9.16 revealed two insert sizes, and six short and long inserts were sequenced. Six clones were sequenced from a control pool of ten unstimulated cells.

Genomic DNA and PCR

Genomic LR-PCR (Expand Long Template, Roche) was carried out from 400 ng DNA, isolated from larval carcass or 0.25 ml blood pelleted for 5 min at 50 g. VLR diversity regions were amplified using LRR.F1 + LRR.R1 (0.5–0.8 kb; animals 10, 12). Mature VLRs were amplified using Slit.F + LRR_C.R2 or LRR_N.F2 + LRR_C.R1 (~1 or ~1.7 kb respectively; animals 10, 11). Amplification across the *gVLR* was with *gVLR.F1* + *gVLR.R1* (animals 10, 13), and the 8-kb amplicon was cloned in pCR-XL (Invitrogen). Pools of 100 erythrocytes or blood lymphocytes were sorted into 0.2 ml TE, 0.1% SDS, 5 µl β-mercaptoethanol and 5 µg proteinase K (Invitrogen). After 1 h at 37 °C and precipitation *gVLR* was amplified by two rounds of nested PCR, first Slit.F + LRR_C.R2 then LRR_N.F2 + LRR_C.R1.

Virtual northern and genome blots

Virtual northern blot was prepared according to the SMART manual. Twenty-cycle amplified cDNA was from unstimulated larvae (tail, liver and sorted lymphocytes from blood, typhlosole and kidneys) and blood cells sorted from four times immunostimulated larvae (small and large lymphocytes, myeloid cells and erythrocytes).

For genome blots, 10 µg larval DNA per lane (animal 10, 12, 13) was digested with *Bam*HI, *Eco*RI and *Hind*III (Roche). For the pulse-field blot (CHEF DRIII, Bio-Rad), erythrocytes from ten larvae were embedded in agarose⁴¹ and 20 µg DNA per lane was digested with *Asc*I, *Fse*I, *Nor*I, *Pac*I, *Pme*I and *Sfi*I (New England Biolabs).

³²P-labelled probes were: VLR N terminus (196 bp) amplified with Slit.F + Slit.R, and C terminus (208 bp) amplified with LRR_C.F1 + LRR_C.R1; GAPDH (314 bp) was amplified with GAPDH.F + GAPDH.R.

PAC clones

The arrayed sea lamprey PAC library in pCYPAC6 (AF133437) was constructed from erythrocytes of one Lake Michigan adult using partial *Mbo*I digests⁴¹. The 6 × 10⁴ clones had ~65-kb inserts with 1–2-fold genome coverage. The library was screened with N-terminal and C-terminal probes; five PACs hybridized with both probes (2, 3, 15–17) and six others hybridized only with the N terminus (4, 9, 14, 35, 42, 43). *gVLR* was amplified from PACs 2, 3, 15–17 using Slit.F + LRR_C.R2. The 14-kb amplicons revealed

two sets of *Bam*HI patterns: PACs 2, 3 and 15–17. PACs 3, 4, 16 were sequenced at McGill University.

VLR GPI anchor

A VLR insert, LRRNT to stop codon, was amplified using Dis_LRR.F + Dis_LRR.R1 (Expand High Fidelity) and fused to Igκ signal peptide and haemagglutinin (HA) epitope in pDisplay (Invitrogen). Transfectants in BW1547 cells, the GPI-anchored VLR, or mFcγRIIb as control, were treated with 1 unit ml⁻¹ bacterial GPI-specific phospholipase C (Sigma) for 45 min at 30 °C. Epitope-tagged proteins were stained with anti-HA-tag monoclonal antibody 12CA5 (Roche) for FACS analysis.

Sequence analysis

A genetic distance dendrogram was generated with MEGA 2.1 UPGMA⁴². GPI-anchor site was predicted by means of <http://129.194.185.165/dgpi/>. SWISS-MODEL²⁴ VLR three-dimensional prediction was made using http://cubic.bioc.columbia.edu/predictprotein/submit_meta.html. Residues 22–319 from clone 12.26 were threaded on crystal coordinates of CD42α (1m10.pdb) and NOGO-66 receptor (1p8t.pdb).

Received 17 March; accepted 11 June 2004; doi:10.1038/nature02740.

1. Litman, G. W., Anderson, M. K. & Rast, J. P. Evolution of antigen binding receptors. *Annu. Rev. Immunol.* **17**, 109–147 (1999).
2. Flajnik, M. F. & Kasahara, M. Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. *Immunity* **15**, 351–362 (2001).
3. Flajnik, M. F. Comparative analyses of immunoglobulin genes: surprises and portents. *Nature Rev. Immunol.* **2**, 688–698 (2002).
4. Forey, P. L. & Janvier, P. Agnathans and the origin of jawed vertebrates. *Nature* **361**, 129–134 (1993).
5. Finstad, J. & Good, R. A. The evolution of the immune response. III. Immunologic responses in the lamprey. *J. Exp. Med.* **120**, 1151–1167 (1964).
6. Marchalonis, J. J. & Edelman, G. M. Phylogenetic origins of antibody structure. 3. Antibodies in the primary immune response of the sea lamprey, *Petromyzon marinus*. *J. Exp. Med.* **127**, 891–914 (1968).
7. Litman, G. W. *et al.* The evolution of the immune response. VIII. Structural studies of the lamprey immunoglobulin. *J. Immunol.* **105**, 1278–1285 (1970).
8. Pollara, B., Litman, G. W., Finstad, J., Howell, J. & Good, R. A. The evolution of the immune response. VII. Antibody to human “O” cells and properties of the immunoglobulin in lamprey. *J. Immunol.* **105**, 738–745 (1970).
9. Good, R. A., Finstad, J. & Litman, G. W. in *The Biology of lampreys II: Immunology* (eds Hardisty, M. V. & Potter, I. C.) 405–432 (Academic, London, 1972).
10. Hagen, M., Filosa, M. F. & Youson, J. H. The immune response in adult sea lamprey (*Petromyzon marinus* L.): the effect of temperature. *Comp. Biochem. Physiol.* **82**, 207–210 (1985).
11. Perey, D. Y., Finstad, J., Pollara, B. & Good, R. A. Evolution of the immune response. VI. First and second set skin homograft rejections in primitive fishes. *Lab. Invest.* **19**, 591–597 (1968).
12. Cooper, A. J. in *Ammocoete Lymphoid Cell Populations In vitro. 4th Leukocyte Culture Conference* (ed. McIntyre, O. R.) 137–147 (Appleton Century-Crofts, New York, 1971).
13. Piavis, G. W. & Hiatt, J. L. Blood cell lineage in the sea lamprey *Petromyzon marinus* (Pisces: Petromyzontidae). *Copeia* **4**, 722–728 (1971).
14. Fujii, T. Electron microscopy of the leukocytes of the typhlosome in ammocoetes, with special attention to the antibody-producing cells. *J. Morphol.* **173**, 87–100 (1982).
15. Ardavin, C. F. & Zapata, A. Ultrastructure and changes during metamorphosis of the lympho-hemopoietic tissue of the larval anadromous sea lamprey *Petromyzon marinus*. *Dev. Comp. Immunol.* **11**, 79–93 (1987).
16. Mayer, W. E. *et al.* Isolation and characterization of lymphocyte-like cells from a lamprey. *Proc. Natl Acad. Sci. USA* **99**, 14350–14355 (2002).
17. Mayer, W. E., O’Huigin, C., Tichy, H., Terzic, J. & Saraga-Babic, M. Identification of two Ikaros-like transcription factors in lamprey. *Scand. J. Immunol.* **55**, 162–170 (2002).
18. Haire, R. N., Miracle, A. L., Rast, J. P. & Litman, G. W. Members of the Ikaros gene family are present in early representative vertebrates. *J. Immunol.* **165**, 306–312 (2000).
19. Shintani, S. *et al.* Do lampreys have lymphocytes? The Spi evidence. *Proc. Natl Acad. Sci. USA* **97**, 7417–7422 (2000).
20. Anderson, M. K., Sun, X., Miracle, A. L., Litman, G. W. & Rothenberg, E. V. Evolution of hematopoiesis: three members of the PU.1 transcription factor family in a cartilaginous fish, *Raja eglanteria*. *Proc. Natl Acad. Sci. USA* **98**, 553–558 (2001).
21. Azumi, K. *et al.* Genomic analysis of immunity in a urochordate and the emergence of the vertebrate immune system: “waiting for Godot”. *Immunogenetics* **55**, 570–581 (2003).
22. Uinuk-Ool, T. S. *et al.* Lamprey lymphocyte-like cells express homologs of genes involved in immunologically relevant activities of mammalian lymphocytes. *Proc. Natl Acad. Sci. USA* **99**, 14356–14361 (2002).
23. Uinuk-Ool, T. S. *et al.* Identification and characterization of a TAP-family gene in the lamprey. *Immunogenetics* **55**, 38–48 (2003).
24. Schwede, T., Diemand, A., Guex, N. & Peitsch, M. V. Protein structure computing in the genomic era. *Res. Microbiol.* **151**, 107–112 (2000).
25. Bell, J. K. *et al.* Leucine-rich repeats and pathogen recognition in Toll-like receptors. *Trends Immunol.* **24**, 528–533 (2003).
26. Bryan, M. B., Libants, S. V., Warrillow, J. A., Li, W. & Scribner, K. T. Polymorphic microsatellite markers for the landlocked sea lamprey, *Petromyzon marinus*. *Conserv. Genet.* **4**, 113–116 (2003).
27. Beutler, B. Innate immunity: an overview. *Mol. Immunol.* **40**, 845–859 (2004).
28. Landmann, R., Müller, B. & Zimmerli, W. CD14, new aspects of ligand and signal diversity. *Microbes Infect.* **2**, 295–304 (2000).
29. Chamailard, M., Girardin, S. E., Viala, J. & Philpott, D. J. Nods, Nalps and Naip: intracellular regulators of bacterial-induced inflammation. *Cell. Microbiol.* **5**, 581–592 (2003).
30. Jones, D. A. & Takemoto, D. Plant innate immunity—direct and indirect recognition of general and specific pathogen-associated molecules. *Curr. Opin. Immunol.* **16**, 48–62 (2004).
31. Ikezawa, H. Glycosylphosphatidylinositol (GPI)-anchored proteins. *Biol. Pharm. Bull.* **25**, 409–417 (2002).
32. Kaufman, J. The origins of the adaptive immune system: whatever next? *Nature Immunol.* **3**, 1124–1125 (2002).
33. Thompson, C. B. & Nieman, P. E. Somatic diversification of the chicken immunoglobulin light chain gene is limited to the rearranged variable gene segment. *Cell* **48**, 369–378 (1987).
34. Reynaud, C. A., Anquez, V., Grimal, H. & Weill, J. C. A hyper conversion mechanism generates the chicken light chain preimmune repertoire. *Cell* **48**, 379–388 (1987).
35. Hamrick, T. S., Dempsey, J. A., Cohen, M. S. & Cannon, J. G. Antigenic variation of gonococcal pilin expression *in vivo*: analysis of the strain FA1090 pilin repertoire and identification of the *pilS* gene copies recombining with *pilE* during experimental human infection. *Microbiol.* **147**, 839–849 (2001).
36. Wang, D., Botkin, D. J. & Norris, S. J. Characterization of the *vls* antigenic variation loci of the Lyme disease spirochaetes *Borrelia garinii* Ip90 and *Borrelia afzelii* ACAL. *Mol. Microbiol.* **47**, 1407–1417 (2003).
37. Berriman, M. *et al.* The architecture of variant surface glycoprotein gene expression sites in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **122**, 131–140 (2002).
38. Duffy, M. F., Reeder, J. C. & Brown, G. V. Regulation of antigenic variation in *Plasmodium falciparum*: censoring freedom of expression? *Trends Parasitol.* **19**, 121–124 (2003).
39. Nash, T. E. Surface antigenic variation in *Giardia lamblia*. *Mol. Microbiol.* **45**, 585–590 (2002).
40. Schluter, S. F., Bernstein, R. M., Bernstein, H. & Marchalonis, J. J. ‘Big Bang’ emergence of the combinatorial immune system. *Dev. Comp. Immunol.* **23**, 107–111 (1999).
41. Amemiya, C. T., Ota, T. & Litman, G. W. in *Nonmammalian Genomic Analysis: A Practical Guide. Construction of P1 Artificial Chromosome (PAC) Libraries from Lower Vertebrates* (eds Birren, B. & Lai, E.) 223–256 (Academic, New York, 1996).
42. Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**, 1244–1245 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. Bergstedt and W. Swink of the Hammond Bay Biological Station for assistance with lamprey biology and supply; G. Hines and the UAB Animal Resources Program staff for attending the lampreys; Y. Kubagawa for VLR sequencing; P. Burrows, H. Schroeder and E. Sorscher for discussion; and A. Brookshire for manuscript preparation. The work was supported by grants from the National Institutes of Health and the National Science Foundation. M.D.C. is an Investigator of the Howard Hughes Medical Institute and Z.P. was a recipient of the Cottrell postdoctoral award.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.D.C. (max.cooper@ccc.uab.edu). All sequences reported here are deposited in GenBank under accession numbers AY577941–AY578059 and CK988414–CK988652.

A high abundance of massive galaxies 3–6 billion years after the Big Bang

Karl Glazebrook¹, Roberto G. Abraham², Patrick J. McCarthy³, Sandra Savaglio¹, Hsiao-Wen Chen⁴, David Crampton⁵, Rick Murowinski³, Inger Jørgensen⁶, Kathy Roth⁶, Isobel Hook⁷, Ronald O. Marzke⁸ & R. G. Carlberg²

¹Department of Physics & Astronomy, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218-2686, USA

²Department of Astronomy & Astrophysics, University of Toronto, 60 St George Street, Toronto, Ontario M5S 3H8, Canada

³Observatories of the Carnegie Institute of Washington, Santa Barbara Street, Pasadena, California 9110, USA

⁴Center for Space Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307, USA

⁵Herzberg Institute of Astrophysics, National Research Council, 5071 West Saanich Road, Victoria, British Columbia, V9E 2E7, Canada

⁶Gemini Observatory, Hilo, Hawaii 96720, USA

⁷Department of Astrophysics, Nuclear & Astrophysics Laboratory, Oxford University, Keble Road, Oxford OX1 3RH, UK

⁸Department of Physics and Astronomy, San Francisco State University, 1600 Holloway Avenue, San Francisco, California 94132, USA

Hierarchical galaxy formation is the model whereby massive galaxies form from an assembly of smaller units¹. The most massive objects therefore form last. The model succeeds in describing the clustering of galaxies², but the evolutionary history of massive galaxies, as revealed by their visible stars and gas, is not accurately predicted. Near-infrared observations (which allow us to measure the stellar masses of high-redshift galaxies³) and deep multi-colour images indicate that a large fraction of the stars in massive galaxies form in the first 5 Gyr (refs 4–7), but uncertainties remain owing to the lack of spectra to confirm the redshifts (which are estimated from the colours) and the role of obscuration by dust. Here we report the results of a spectroscopic redshift survey that probes the most massive and quiescent galaxies back to an era only 3 Gyr after the Big Bang. We find that at least two-thirds of massive galaxies have appeared since this era, but also that a significant fraction of them are already in place in the early Universe.

Star-formation rates in the high-redshift ($z > 1$) Universe have been probed by measuring rest-frame ultraviolet (UV) emission⁸ of galaxies. The UV is dominated by newly formed massive stars but can be highly obscured⁹. Moreover, star-formation in galaxies is a stochastic phenomenon and correlates poorly with mass: the most massive galaxies in the local Universe are giant ellipticals and they have very weak UV emission¹⁰, which makes them difficult to detect at $z > 1$. Direct determination of dynamical masses of galaxies requires spatially resolved velocity measurements of galaxies—this has been done out to $z \approx 1$ (refs 11,12), but such observations are very challenging at higher redshifts owing to signal-to-noise limitations.

An alternative approach is to infer the total mass in stars from red light, which traces accumulated stellar populations. Studies³ have shown that stellar mass correlates extremely well with dynamical mass out to $z = 1$. Stellar mass evolution can be predicted by using galaxy formation models based on cosmological numerical simulations augmented with analytical star-formation recipes¹³, which adopt simplified prescriptions for various heating and cooling processes in the interstellar medium^{13,14}.

Deep near-infrared data provide a window on stellar masses of galaxies at high redshift. This light is dominated by the old, evolved stellar populations in galaxies, and is little affected by transient star-

formation¹⁵ or dust obscuration. To a good approximation, the total K-band light traces the accumulation of stellar mass; equivalently, one can say that the stellar mass-to-light ratio (M/L_K) is nearly constant, with little dependence on the previous star-formation history (SFH). In fact, rest-frame M/L_K varies only by a factor of two between extremely young and extremely old galaxy stellar populations: in contrast, M/L_B can vary by more than a factor of ten¹⁶, and M/L_{UV} varies tremendously (reaching as low as 1% of the solar value) because the UV light is dominated by the instantaneous star-formation rate (SFR) rather than the stellar mass. Even at redshifts up to $z = 2$, the observed-frame K-band probes the rest-frame R-band which is still dominated by old stellar populations. The M/L_R variation is still less than a factor of four, partly due to the youth of the Universe at that redshift.

We use the Gemini Deep Deep Survey (GDDS)¹⁷ to define a new sample of 150 galaxies with $K < 20.6$ mag and $0.8 < z < 2$ located in four independent 30 arcmin² fields. The spectroscopic identification completeness is 89%. The determination of stellar masses for each galaxy follows standard multi-colour stellar population fitting techniques; this approach is fairly general and robust, and is detailed in Methods. We find that the K-band light traces the stellar mass quite well: evolutionary changes in galaxy numbers with redshift are much more important than changes in colour. The findings presented below are simply driven by the presence of numerous $K \approx 20$ galaxies at $z > 1.5$, which must be massive objects.

Figures 1 and 2 illustrate the nature of the most massive galaxies. They are found by the GDDS to $z = 2$ at $K \approx 20$, even when they are very red ($I - K > 4$). The red galaxies are predominantly red owing to old stellar populations (these are GDDS spectral classes¹⁷ '001' and '011', showing photospheric features from evolved stars) and not dust reddening of young lower-mass galaxies. The red, old

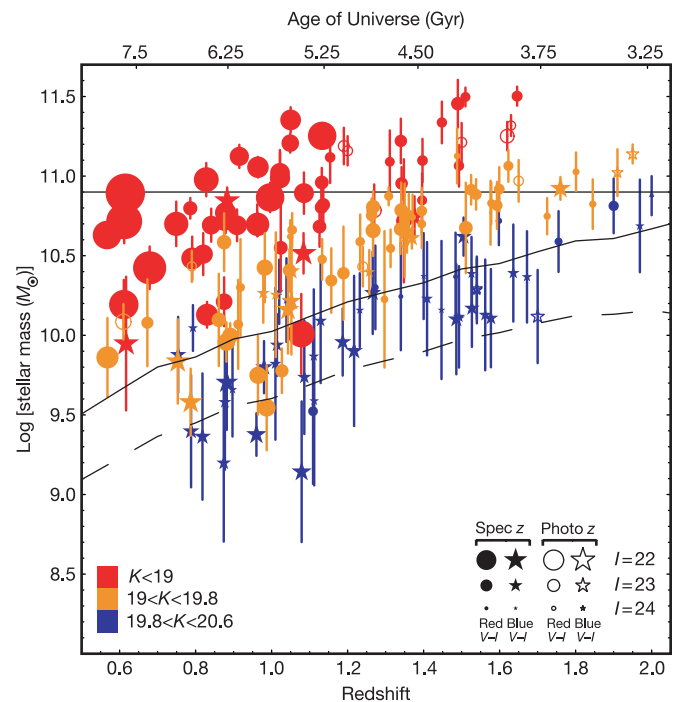


Figure 1 Stellar mass–redshift distribution for our galaxies. The 1σ error bars come from the Monte Carlo mass fitting. Symbol colours code the observed K-band magnitude (see key on left). Solid/open symbol shapes denote spectroscopic/photometric redshifts respectively. Circle/star symbols denote objects respectively redder/bluer in $V - I$ than a model Sbc galaxy template²⁰. Symbol size is keyed to the I-band magnitude. The horizontal line denotes the characteristic Schechter mass scale in the local Universe²⁰. The solid curve shows how the K-band flux limit translates into a mass completeness limit for a maximally old simple stellar population (SSP). The dashed curve shows an example mass limit for bluer objects (SFR = const. model).

galaxies make a large contribution (30%) to the stellar mass in the Universe in the redshift range $1.2 < z < 1.8$. Also at $z > 1$ there are a number of blue galaxies despite this being a K-selected sample. These correspond to massive star-forming galaxies which have high metal abundances¹⁸.

The cumulative stellar mass density per unit volume in each redshift bin down to various mass thresholds (Fig. 3; Table 1) is computed following the standard maximum volume V/V_{max} formalism (this corrects for the smaller redshift ranges covered by fainter objects¹⁹) for a $K < 20.6$ limit and weighting by the sampling. We use K-corrections from our individual spectral energy distribution (SED) fits, but the results do not depend strongly on the details of the K-correction. Although a K-selected sample is a good proxy for a mass-selected sample, it is still necessary to consider incompleteness as a function of M/L . To do this, we compute the maximum possible M/L at each redshift for a model galaxy as old as the Universe which formed all its stars at once (a 'simple stellar population', or SSP). This M/L is converted to a mass limit (via the K-band flux limit and the K-correction) and is shown in Fig. 1. We are complete above this mass limit; bluer objects can be seen below this limit. Bins that might miss high M/L objects are plotted as lower limits in Fig. 3. The error bars are calculated from shot noise on the number of galaxies in each bin. These do not include the effects of large-scale structure, but we believe these are not significant because our GDDS fields are large, were selected from even larger-area images in regions near average density¹⁷ (that is, neither highly over-dense nor under-dense) and because our colour-dependent weights normalize to the full imaging area (554.7 arcmin^2) which would counteract the additional clustering of red objects. Finally, we have performed the check of splitting the sample by different fields (different independent sight-lines). The

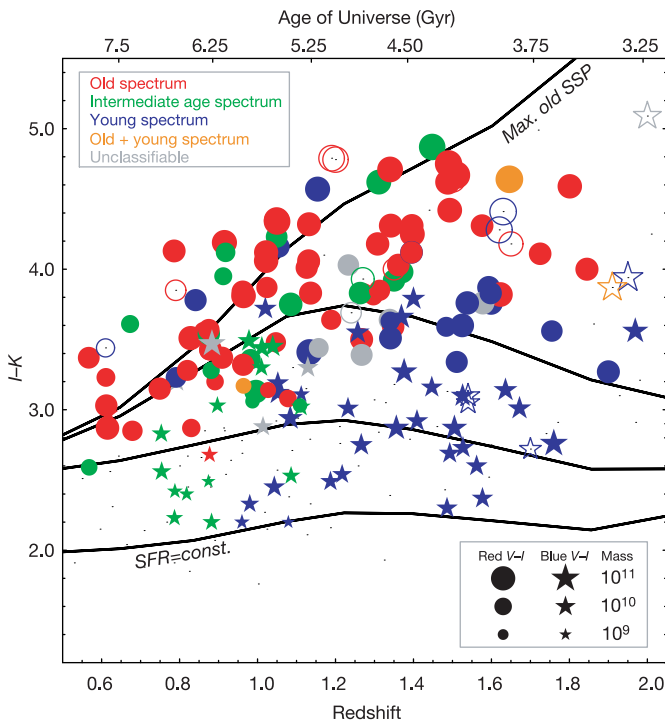


Figure 2 Colour-redshift distribution for our galaxies. Observed frame $I - K$ colour is plotted with symbol size keyed to mass, shape to optical $V - I$ colour (as in Fig. 1) and symbol colour keyed to the spectral classification (see key at top). Solid/open symbols denote spectroscopic/photometric redshifts. Model tracks (solid lines) are shown ranging from a maximally old SSP through to $\text{SFR} = \text{const.}$ for synthetic solar metallicity galaxies which form at $z = 10$. (Note the points red-ward of the old track at $z = 0.6$ may be artificially reddened by observational aperture/metallicity effects; this is insignificant for $z > 0.8$.)

same general results are found for these, albeit with larger errors. We have also assessed the effect of the mass fitting errors on the mass densities with our Monte Carlo methods; this is not a significant source of error. In every bin, galaxies with spectroscopic redshifts dominate the mass budget, except for the $1.6 < z < 2$ and $M > 10^{10.8} M_{\odot}$ bin, where the spectroscopic completeness is only 50% and we augment our spectroscopy with photometric redshifts. Analysing the sub-sample with only spectroscopic redshifts results in no significant changes to any bin except for this one (which is thus 0.3 dex lower); the essential scientific result is unchanged.

There is a clear decline with redshift in stellar mass locked up in the most massive galaxies. This trend is in accord with ideas of gradual assembly. However, it declines only slowly towards high redshift and, surprisingly, the massive galaxies do not decline more rapidly than the whole population. We note that galaxies with $M > 10^{10.8} M_{\odot}$ are brighter than our flux limits (for any possible M/L value) throughout our redshift range; a regression on these gives an acceptable fit for $\rho \propto (1 + z)^{-1.7 \pm 1.6}$. At $z = 1$ the mass densities for the $M > 10^{10.8} M_{\odot}$ sample are $38 (\pm 18)\%$ of their local value²⁰; at $z = 1.8$ this becomes $16 (\pm 6)\%$. These results for the most massive galaxies are consistent with previous Hubble Deep Field South photometric redshift determinations^{4,5}, but inconsistent with the more rapid decline (a factor of ~ 6 over $0 < z < 1$) found with photometric redshifts by the Munich Near-Infrared Cluster Survey⁶ (however we note that their SED fitting forces maximally old galaxies at all redshifts, which could be problematic). Overlaid on Fig. 3 (shaded region) is the range of estimates for the growth of total stellar mass based on the integral of the observed rest-UV derived SFR- z relationship. (These are based on the analytic fit²⁰ of points from Fig. 9 of ref. 9, both with and without extinction correction, integrated using PEGASE.2²¹.) Our uppermost points represent lower limits on the total stellar mass density, and we confirm previous findings that an extinction correction is essential for UV SFR estimates to be consistent with stellar mass measurements^{4,20}.

Theoretical models of galaxy formation convert gas into stars in dark matter haloes using semi-empirical recipes, and must satisfy three key observables: the first is the distribution of local galaxy luminosities, the second is the distribution of galaxy colours and the third is the abundance of massive galaxies at high redshift. The last of these has until now been the most difficult to measure from observations. In Fig. 3 we plot the abundance predictions of the

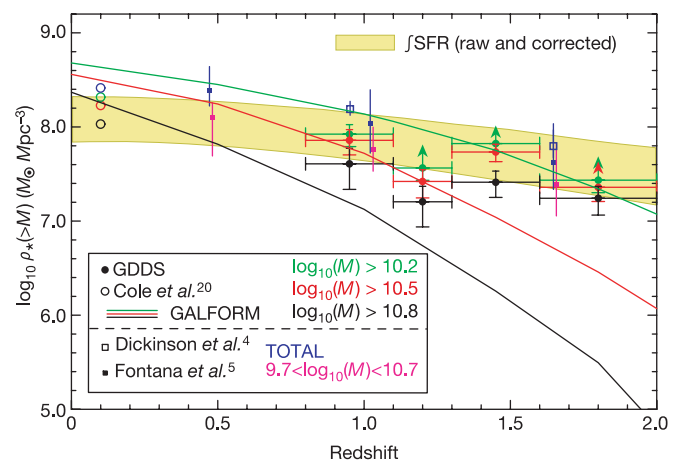


Figure 3 Mass density in stars versus redshift. Values from our spectroscopic sample are compared with previous estimates which are all based on photometric redshifts^{4,5} (except for the local $z = 0.1$ point²⁰). We plot the cumulative mass density of galaxies (converted to our IMF choice; error bars are 1σ more massive than a given mass threshold (see key for threshold colour coding)). Theoretical 'GALFORM' models are also plotted (see text). The shaded region shows the result of integrating the Universal UV-derived star-formation history with and without a dust extinction correction.

'GALFORM' models^{14,22} which satisfy the first two constraints. It is evident in this model that massive galaxies disappear much more rapidly than we see in our data; this is because the model stellar mass build-up traces the merging of cold dark matter haloes. In particular, GALFORM has a strong dependence of evolutionary rate on mass which is not seen in our data. A similar theoretical comparison was made on the Hubble Deep South photometric data with similar conclusions⁵.

We note that our measured abundance of massive galaxies at $z \approx 2$ does not violate the boundary constraints of cold dark matter models, by which we mean taking the predicted abundance of $z = 2$ massive dark matter haloes²³ and scaling by the cosmological baryon / dark matter density² ($\Omega_b/\Omega_m = 0.17$). This gives a predicted baryonic mass density of $\sim 10^8 M_\odot \text{Mpc}^{-3}$ in haloes with baryonic masses $> 10^{11} M_\odot$, a factor of ten above our measurements. Thus the models can match our densities if only $\sim 10\%$ of the baryons in massive haloes are converted to stars by $z = 2$. However, we note that in the Universe today, the stellar density²⁰ $\Omega_* = 0.004$ gives $\Omega_*/\Omega_b = 0.1$; this would imply that massive haloes at $z = 2$ have managed to convert baryons into stars with the same overall efficiency as the average Universe had achieved by $z = 0$. Massive haloes must have much greater star-formation efficiencies at earlier times.

Models have been proposed^{13,24} which adjust the star-formation histories in this way, and which would be more in accord with our findings of a high abundance of massive galaxies at early times. However, they fail to match existing data on galaxy colours^{25,26} or the galaxy luminosity function¹³, so we defer detailed comparison to future work.

What is observationally clear is that we have measured directly the abundance of massive galaxies out to $z = 2$ for the first time from a highly complete, deep and relatively wide area survey with secure redshifts and spectroscopic classifications for galaxies. We find (with 99% confidence) that at least two-thirds of the mass in these objects has formed since $z = 1.8$. However, we do not find a more rapid evolution of the giant population compared to the Universal average, as would be expected if stellar mass grew simply proportionally to dark matter assembly. At $z \lesssim 2$ we find that much of the mass is in galaxies with old stellar populations; these objects are plausible precursors of modern massive elliptical galaxies. This is broadly consistent with developing ideas that models of galaxy formation have to be tuned towards earlier star-formation in high-mass haloes. The abundance of massive galaxies at $z \lesssim 2$ is now firmly established, and will provide the vital missing leg in the tripod of key observations required to understand how galaxies form. $\square$

Table 1 Stellar mass density measurements in the Universe

$\log M_{\text{lim}}$	BG03 IMF				Kennicutt IMF			
	z range	$\log \rho$	$\log \rho_{\text{lo}}$	$\log \rho_{\text{hi}}$	$\log \rho$	$\log \rho_{\text{lo}}$	$\log \rho_{\text{hi}}$	Complete?
10.2	0.8–1.1	7.92	7.79	8.02	7.82	7.67	7.92	Y
10.2	1.1–1.3	7.56	7.43	7.66	7.47	7.34	7.58	N
10.2	1.3–1.6	7.82	7.73	7.90	7.71	7.62	7.79	N
10.2	1.6–2.0	7.43	7.30	7.54	7.34	7.21	7.44	N
10.5	0.8–1.1	7.86	7.70	7.97	7.72	7.54	7.84	Y
10.5	1.1–1.3	7.42	7.25	7.55	7.34	7.17	7.47	Y
10.5	1.3–1.6	7.73	7.63	7.82	7.62	7.51	7.71	Y
10.5	1.6–2.0	7.36	7.21	7.47	7.26	7.11	7.37	N
10.8	0.8–1.1	7.61	7.34	7.77	7.46	7.13	7.64	Y
10.8	1.1–1.3	7.21	6.94	7.37	6.92	6.49	7.14	Y
10.8	1.3–1.6	7.41	7.25	7.53	7.24	7.04	7.38	Y
10.8	1.6–2.0	7.24	7.06	7.37	6.96	6.74	7.11	Y

Measurements are shown as a function of galaxy mass threshold and redshift. ρ is the stellar mass density in galaxies with stellar masses $> M_{\text{lim}}$ in $M_\odot \text{Mpc}^{-3}$. The values of ρ_{hi} and ρ_{10} are the 1σ upper and lower limits from counting statistics. BG03 IMF mass thresholds and mass densities can be multiplied by 1.82 to convert to the Salpeter IMF.

Methods

Stellar mass fitting

We adopt a cosmology² of $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$, $H_0 = 70 \text{ km s}^{-1} \text{Mpc}^{-1}$ and use Vega magnitudes. For the mass-function analysis, we use GDDS galaxies with $K < 20.6$ and $0.8 < z < 2$ (150 galaxies; 89% spectroscopic completeness based on ref. 17 identification confidence classes ≥ 2 ; galaxies without spectroscopic redshifts are assigned to their photometric redshifts). To account for the higher priority given to red galaxies in the slit-mask design we use the 'sampling weights' from ref. 17, which are the selection probabilities for spectroscopy as a function of $I - K$ and K .

For each galaxy we derive the most likely stellar mass and a range of uncertainty. These are derived by evaluating against model optical-IR SEDs to determine the M/L_K ratio, and hence the mass. We note that a variety of approaches to accomplish this have been described in the literature which vary in the level of detail in which they treat star formation. For example, ref. 20 used a simple set of monotonic star-formation histories (SFH) with a varying e-folding timescale and a fixed dust law. A potential problem is posed by the fact that real galaxies have more complex SFHs; for example, a recent star-burst can make an old galaxy temporarily bluer and lead to an underestimate of M/L using this method. One approach to account for this effect is to introduce a second young SED component⁵ superimposed on the old population; this can be computationally expensive depending on the amount of freedom allowed for the second component.

In a spectroscopic sample there is no degeneracy between SED model fitting and photometric redshift (which is also based on SED fitting). We adopt two-component modelling (using PEGASE.2²¹ to calculate spectra) in order to be able to assess the possible biases due to starbursts on the calculated masses, and allow a range of dust extinction ($0 \leq A_V \leq 2 \text{ mag}$) and metallicity ($0.0004 \leq Z \leq 0.02$). The primary component is modelled using a star-formation rate $\text{SFR} \propto \exp(-t/\tau)$ with $\tau = 0.1, 0.2, 0.5, 1, 2, 4, 8$ and 500 Gyr (the first approximates an instantaneous starburst and the last a constant SFR). The secondary component is a starburst, modelled with a $\tau = 0.1$ Gyr exponential, which can occur at any time and have a mass between 10^{-4} and twice that of the primary component. The model age is constrained to be less than that of the Universe at the appropriate redshift, but any formation epoch is allowed. Like all studies of high redshift star-formation, we must assume a universal initial mass function (IMF). We primarily use the BG03 IMF²⁷ which fits local cosmic luminosity densities well; it has a similar high-mass slope to the classical Salpeter IMF²⁸ but with a more realistic break at $0.5 M_\odot$. Stars with masses below the break never contribute significantly to optical/IR light, so the overall effect is to re-scale the total stellar masses to more reasonable values. Similarly we re-scale literature Salpeter based numbers^{4,5,20} to the BG03 IMF using $M_{\text{SP}} = 1.82 M_{\text{BG}}$. Our results are robust for any reasonable choice of IMF with a similar slope; to illustrate this, we have also calculated masses using the popular Kennicutt IMF²⁹.

Our approach is to find via exhaustive grid search all models consistent with the $VIZ'K$ photometry in the observed frame of each galaxy. This colour set is available for all galaxies, and covers rest-frame UV through near IR. We do not include the actual spectra (apart from the redshift information) in the fits because of variable quality and signal-to-noise ratio, but we find the spectral classes are broadly consistent with the best photometric SED fits. The full distribution function of allowed masses were calculated by Monte Carlo re-sampling of the photometric errors. The final masses and error bars represent the mean and standard deviation of this full distribution function. Typically, we find the masses are fitted to ± 0.17 dex in the $K < 20.6$ sample. The stellar masses are very robust against the details of the fitting. Using the Monte Carlo machinery to investigate the effect of different assumptions about metallicity, dust and bursts, we find that the largest effect is due to bursts. If we disallowed bursts, then the masses typically decrease by only 0.2 dex. Finally, we note that the variation in M/L_K over the range $1 < z < 2$ is constrained by the age of the Universe (6–3 Gyr); typically the maximum range in our sample is a factor of three.

Received 4 January; accepted 24 May 2004; doi:10.1038/nature02667.

- Blumenthal, G. R., Faber, S. M., Primack, J. R. & Rees, M. J. Formation of galaxies and large-scale structure with cold dark matter. *Nature* **311**, 517–525 (1984).
- Spergel, D. N. *et al.* First-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Determination of cosmological parameters. *Astrophys. J. Suppl.* **148**, 175–194 (2003).
- Brinchmann, J. & Ellis, R. S. The mass assembly and star formation characteristics of field galaxies of known morphology. *Astrophys. J.* **536**, L77–L80 (2000).
- Dickinson, M., Papovich, C., Ferguson, H. C. & Budavári, T. The evolution of the global stellar mass density at $0 < z < 3$. *Astrophys. J.* **587**, 25–40 (2003).
- Fontana, A. *et al.* The assembly of massive galaxies from near-infrared observations of the Hubble Deep Field-South. *Astrophys. J.* **594**, L9–L12 (2003).
- Drory, N. *et al.* The Munich Near-Infrared Cluster Survey: Number density evolution of massive field galaxies to $z \sim 1.2$ as derived from the K-band-selected survey. *Astrophys. J.* **562**, L111–L114 (2001).
- Franx, M. *et al.* A significant population of red, near-infrared-selected high-redshift galaxies. *Astrophys. J.* **587**, L79–L82 (2003).
- Madau, P. *et al.* High-redshift galaxies in the Hubble Deep Field: colour selection and star formation history to $z \sim 4$. *Mon. Not. R. Astron. Soc.* **283**, 1388–1404 (1996).
- Steidel, C. C., Adelberger, K. L., Giavalisco, M., Dickinson, M. & Pettini, M. Lyman-break galaxies at $z \sim 4$ and the evolution of the ultraviolet luminosity density at high redshift. *Astrophys. J.* **519**, 1–17 (1999).
- Lotz, J. M., Ferguson, H. C. & Bohlin, R. C. Mid-ultraviolet determination of elliptical galaxy abundances and ages. *Astrophys. J.* **532**, 830–844 (2000).
- Vogt, N. P. *et al.* Optical rotation curves of distant field galaxies: Sub- L^* systems. *Astrophys. J.* **479**, L121–L124 (1997).
- Gebhardt, K. *et al.* The DEEP Groth strip Survey. IX. Evolution of the fundamental plane of field galaxies. *Astrophys. J.* **597**, 239–262 (2003).
- Kauffmann, G., Colberg, J. M., Diaferio, A. & White, S. D. M. Clustering of galaxies in a hierarchical universe - II. Evolution to high redshift. *Mon. Not. R. Astron. Soc.* **307**, 529–536 (1999).

14. Baugh, C. M., Benson, A. J., Cole, S., Frenk, C. S. & Lacey, C. in *The Mass of Galaxies at Low and High Redshift* (eds Bender, R. & Renzini, A.) 91–96 (ESO Astrophysics Symposia, Berlin, Springer, 2003).
15. Rix, H. & Rieke, M. J. Tracing the stellar mass in M51. *Astrophys. J.* **418**, 123–134 (1993).
16. Bell, E. F., McIntosh, D. H., Katz, N. & Weinberg, M. D. The optical and near-infrared properties of galaxies. I. Luminosity and stellar mass functions. *Astrophys. J. Suppl.* **149**, 289–312 (2003).
17. Abraham, R. G. *et al.* The Gemini Deep Deep Survey: I. Introduction to the survey, catalogs and composite spectra. *Astron. J.* **127**, 2455–2483 (2004).
18. Savaglio, S. *et al.* The Gemini Deep Deep Survey. II. Metals in star-forming galaxies at redshift $1.3 < z < 2$. *Astrophys. J.* **602**, 51–65 (2004).
19. Schmidt, M. Space distribution and luminosity functions of quasi-stellar radio sources. *Astrophys. J.* **151**, 393–409 (1968).
20. Cole, S. *et al.* The 2dF galaxy redshift survey: near-infrared galaxy luminosity functions. *Mon. Not. R. Astron. Soc.* **326**, 255–273 (2001).
21. Fioc, F. & Rocca-Volmerange, B. PEGASE: a UV to NIR spectral evolution model of galaxies. Application to the calibration of bright galaxy counts. *Astron. Astrophys.* **326**, 950–962 (1997).
22. Granato, G. L. *et al.* The infrared side of galaxy formation. I. The local universe in the semianalytical framework. *Astrophys. J.* **542**, 710–730 (2000).
23. Sheth, R. K. & Tormen, G. Large-scale bias and the peak background split. *Mon. Not. R. Astron. Soc.* **308**, 119–126 (1999).
24. Somerville, R. S., Primack, J. R. & Faber, S. M. The nature of high-redshift galaxies. *Mon. Not. R. Astron. Soc.* **320**, 504–528 (2001).
25. Somerville, R. S. *et al.* The redshift distribution of near-infrared-selected galaxies in the Great Observatories Origins Deep Survey as a test of galaxy formation scenarios. *Astrophys. J.* **600**, L135–L138 (2004).
26. Stanford, S. A. *et al.* The evolution of early-type field galaxies selected from a NICMOS map of the Hubble Deep Field North. *Astron. J.* **127**, 131–155 (2004).
27. Baldry, I. K. & Glazebrook, K. Constraints on a universal stellar initial mass function from ultraviolet to near-infrared galaxy luminosity densities. *Astrophys. J.* **593**, 258–271 (2003).
28. Salpeter, E. E. The luminosity function and stellar evolution. *Astrophys. J.* **121**, 161–167 (1955).
29. Kennicutt, R. C. The rate of star formation in normal disk galaxies. *Astrophys. J.* **272**, 54–67 (1983).
30. Coleman, G. D., Wu, C.-C. & Weedman, D. W. Colors and magnitudes predicted for high redshift galaxies. *Astrophys. J. Suppl.* **43** (Suppl.), 393–416 (1980).

Acknowledgements This work is based on observations obtained at the Gemini Observatory, which is operated by AURA under a co-operative agreement with the NSF on behalf of the Gemini partnership: NSF (US), PPARC (UK), NRC (Canada), CONICYT (Chile), ARC (Australia), CNPq (Brazil) and CONICET (Argentina); it is also based on observations made at the Las Campanas Observatory of the Carnegie Institution of Washington. K.G. and S.S. acknowledge funding from the David and Lucille Packard Foundation; S.S. is on leave of absence from the Osservatorio Astronomico di Roma, Italy. R.A. acknowledges funding from NSERC and from the Government of Ontario through a Premier's Research Excellence award. H.-W.C. acknowledges support by NASA through a Hubble Fellowship grant from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.G. (kgb@pha.jhu.edu).

Old galaxies in the young Universe

A. Cimatti¹, E. Daddi², A. Renzini², P. Cassata³, E. Vanzella³, L. Pozzetti⁴, S. Cristiani⁵, A. Fontana⁶, G. Rodighiero³, M. Mignoli⁴ & G. Zamorani⁴

¹INAF - Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, I-50125, Firenze, Italy

²European Southern Observatory, Karl-Schwarzschild-Strasse 2, D-85748, Garching, Germany

³Dipartimento di Astronomia, Università di Padova, Vicolo dell'Osservatorio, 2, I-35122 Padova, Italy

⁴INAF - Osservatorio Astronomico di Bologna, via Ranzani 1, I-40127, Bologna, Italy

⁵INAF - Osservatorio Astronomico di Trieste, via Tiepolo 11, I-34131 Trieste, Italy

⁶INAF - Osservatorio Astronomico di Roma, via dell'Osservatorio 2, Monteporzio, Italy

More than half of all stars in the local Universe are found in massive spheroidal galaxies¹, which are characterized by old stellar populations^{2,3} with little or no current star formation. In present models, such galaxies appear rather late in the history of the Universe as the culmination of a hierarchical merging process, in which larger galaxies are assembled through mergers

of smaller precursor galaxies. But observations have not yet established how, or even when, the massive spheroidals formed^{2,3}, nor if their seemingly sudden appearance when the Universe was about half its present age (at redshift $z \approx 1$) results from a real evolutionary effect (such as a peak of mergers) or from the observational difficulty of identifying them at earlier epochs. Here we report the spectroscopic and morphological identification of four old, fully assembled, massive (10^{11} solar masses) spheroidal galaxies at $1.6 < z < 1.9$, the most distant such objects currently known. The existence of such systems when the Universe was only about one-quarter of its present age shows that the build-up of massive early-type galaxies was much faster in the early Universe than has been expected from theoretical simulations⁴.

In the Λ CDM scenario⁵, galaxies are thought to build up their present-day mass through a continuous assembly driven by the hierarchical merging of dark matter haloes, with the most massive galaxies being the last to form. However, the formation and evolution of massive spheroidal early-type galaxies is still an open question.

Recent results indicate that early-type galaxies are found up to

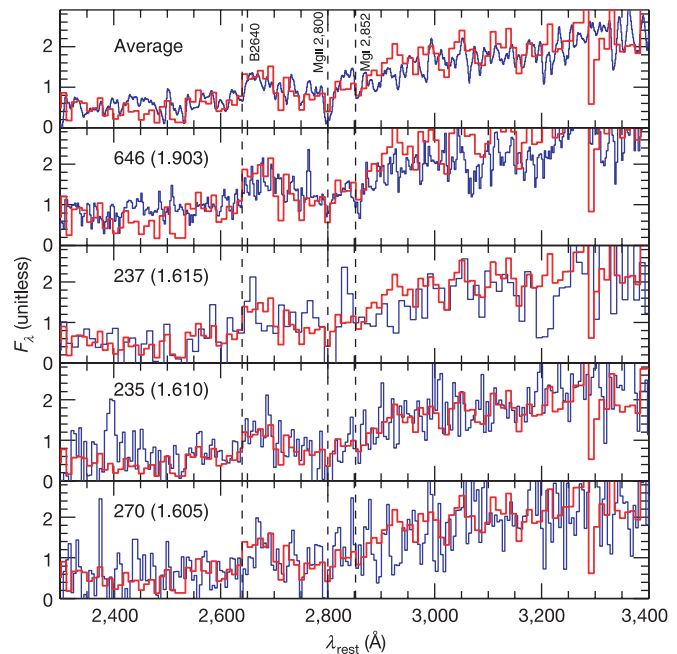


Figure 1 The individual and average spectra of the detected galaxies. From bottom to top: the individual spectra smoothed to a $16 \text{ \AA}$ boxcar ($26 \text{ \AA}$ for ID 237) and the average spectrum of the four old galaxies ($z_{\text{average}} = 1.68$). The red line is the spectrum of the old galaxy LBDS 53w091 ($z = 1.55$) used to search for spectra with a similar continuum shape. Weak features in individual spectra (for example, $\text{MgII } 2,800 \text{ \AA}$ and the $2,640 \text{ \AA}$ continuum break, B2640) become clearly visible in the average spectrum. The object ID 235 has also a weak $[\text{OII}] 3,727 \text{ \AA}$ emission (not shown here). The spectra were obtained with ESO VLT + FORS2, grisms 200I ($R(1'') \approx 400$) (ID 237) and 300I ($R(1'') \approx 600$) (IDs 235, 270, 646), $1.0''$ wide slit and $\leq 1''$ seeing conditions. The integrations times were 3 hours for ID 237, 7.8 hours for IDs 235 and 270, and 15.8 hours for ID 646. For ID 646, the ESO/GOODS public spectrum was co-added to our K20 spectrum (see Supplementary Table 1). 'Dithering' of the targets along the slits was applied to remove efficiently the CCD fringing pattern and the strong OH sky lines in the red. The data reduction was done with the IRAF software package¹⁶. The spectrophotometric calibration of all spectra was achieved and verified by observing several standard stars. The average spectrum, corresponding to 34.4 hours integration time, was obtained by co-adding the individual spectra convolved to the same resolution, scaled to the same arbitrary flux (that is, with each spectrum having the same weight in the co-addition), and assigning wavelength-dependent weights which take into account the noise in the individual spectra due to the OH emission sky lines.

$z \approx 1$ with a number density comparable to that of local luminous E/S0 galaxies^{6,7}, suggesting a slow evolution of their stellar mass density from $z \approx 1$ to the present epoch. The critical question is whether these galaxies do exist in substantial number^{8,9} at earlier epochs, or if they were assembled later^{10,11} as favoured by most renditions of the hierarchical galaxy formation scenario⁴. The problem is complicated also by the difficulty of identifying such galaxies owing to their faintness and, for $z > 1.3$, the lack of strong spectral features in optical spectra, placing them among the most difficult targets even for the largest optical telescopes. For example, whereas star-forming galaxies are now routinely found up to $z \approx 6.6$ (ref. 12), the most distant spectroscopically confirmed old spheroid is still a radio-selected object at $z = 1.552$ discovered a decade ago^{13,14}.

One way of addressing the critical question of massive galaxy formation is to search for the farthest and oldest galaxies with masses comparable to the most massive galaxies in the present-day Universe ($10^{11} - 10^{12} M_{\odot}$ where $M_{\odot}$ is the solar mass), and to use them as the 'fossil' tracers of the most remote events of galaxy formation. As the rest-frame optical–near-infrared luminosity traces the galaxy mass¹⁵,

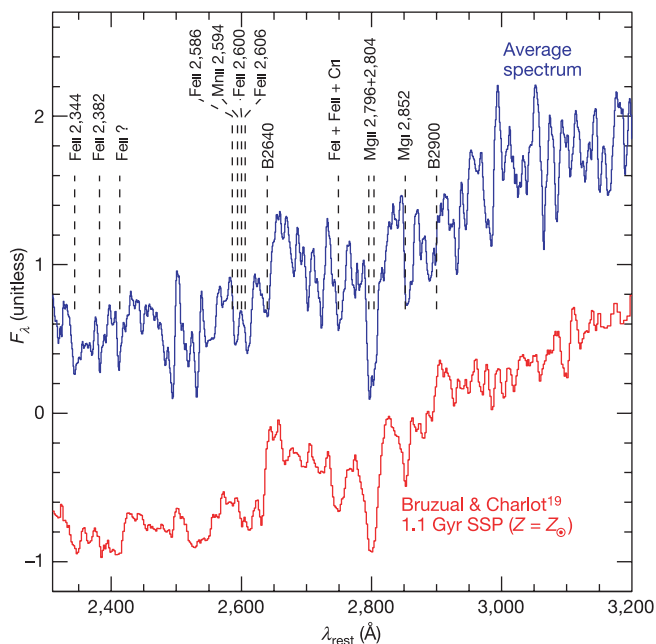


Figure 2 The detailed average spectrum of the detected galaxies. A zoom on the average spectrum (blue) compared with the synthetic spectrum¹⁹ of a 1.1-Gyr-old simple stellar population (SSP) with solar metallicity ($Z = Z_{\odot}$) and Salpeter IMF (red). The observed average spectrum was compared to a library of synthetic SSP template spectra^{19,20} with a range of ages of 0.1–3.0 Gyr with a step of 0.1 Gyr, and with assumed metallicities $Z = 0.4 \times$, $1.0 \times$ and $2.5 \times Z_{\odot}$. The best fit age for each set of synthetic templates was derived through a χ^2 minimization over the rest-frame wavelength range 2,300–3,400 Å. The r.m.s. as a function of wavelength used in the χ^2 procedure was estimated from the average spectrum computing a running mean r.m.s. with a step of 1 Å and a box size of 20 Å, corresponding to about three times the resolution of the observed average spectrum. The median signal-to-noise ratio is ~ 20 per resolution element in the 2,300–3,400 Å range. The wavelength ranges including the strongest real features (that is, absorptions and continuum breaks) were not used in the estimate of the r.m.s. The resulting reduced χ^2 is of the order of unity for the best fit models. In the case of solar metallicity, the ranges of ages acceptable at 95% confidence level are $1.0^{+0.5}_{-0.1}$ Gyr and $1.4^{+0.5}_{-0.4}$ Gyr for SSP models of refs 19 and 20 respectively (see also Fig. 3, top panel). Ages $\sim 50\%$ younger or older are also acceptable for $Z = 2.5Z_{\odot}$ or $Z = 0.4Z_{\odot}$ respectively. The 2,640 Å and 2,900 Å continuum break¹³ amplitudes measured on the average spectrum are $B2640 = 1.8 \pm 0.1$ and $B2900 = 1.2 \pm 0.1$. These values are consistent with the ones expected in SSP models^{19,20} for ages around 1–1.5 Gyr and solar metallicity. For instance, the SSP model spectrum shown here has $B2640 = 1.84$ and $B2900 = 1.27$.

the K_s band ($\lambda \approx 2.2 \mu\text{m}$ in the observer frame) allows a fair selection of galaxies according to their mass up to $z \approx 2$.

Following this approach, we recently conducted the K20 survey¹⁶ with the Very Large Telescope (VLT) of the European Southern Observatory (ESO). Deep optical spectroscopy was obtained for a sample of 546 objects with $K_s < 20$ mag (Vega photometric scale) and extracted from an area of 52 arcmin², including 32 arcmin² within the GOODS–South field¹⁷ (hereafter the GOODS/K20 field). The spectroscopic redshift (z_{spec}) completeness of the K20 survey is 92%, and the available multi-band photometry ($BVRizJHK_s$) allowed us to derive the spectral energy distribution (SED) and photometric redshift (z_{phot}) of each galaxy. The K20 survey spectroscopy was complemented with the ESO/GOODS public spectroscopy (Supplementary Table 1).

The available spectra within the GOODS/K20 field were then used to search for old, massive galaxies at $z > 1.5$. We spectroscopically identified four galaxies with $18 \lesssim K_s \lesssim 19$ and $1.6 \lesssim z_{\text{spec}} \lesssim 1.9$ which have rest-frame mid-ultraviolet (UV) spectra with shapes and continuum breaks compatible with being dominated by old stars and $R - K_s \gtrsim 6$ (the colour expected at $z > 1.5$ for old passively evolving galaxies due to the combination of old stellar populations and k-correction effects⁹). Supplementary Table 1 lists the main galaxy information. The spectrum of each individual object allows a fairly precise determination of the redshift on the basis of absorption features and of the overall spectral shape (Fig. 1).

The co-added average spectrum of the four galaxies (Figs 2 and 3) shows a near-UV continuum shape, breaks and absorption lines that are intermediate between those of a F2 V and a F5 V star¹⁸, and typical of about 1–2-Gyr-old synthetic stellar populations^{19,20}. It is also very similar to the average spectrum of $z \approx 1$ old extremely red objects⁷ (EROs), and slightly bluer than that of the $z \approx 0.5$ SDSS red luminous galaxies²¹ and of the $z = 1.55$ old galaxy LBDS 53w091¹³. However, it is different in shape and slope from the average spectrum of $z \approx 1$ dusty star-forming EROs⁷.

The multi-band photometric SED of each galaxy was successfully

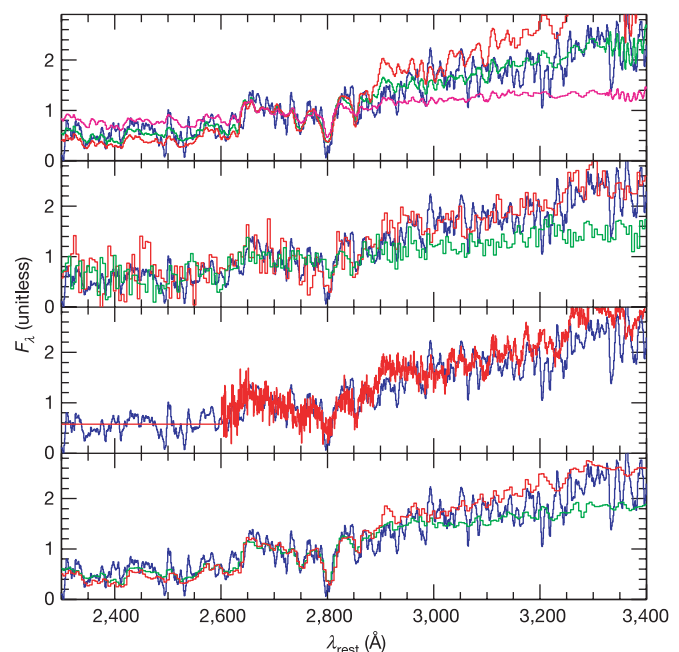


Figure 3 The average spectrum (blue) compared to a set of template spectra. From bottom: F2 V (green) and F5 V (red) stellar spectra¹⁸ with $Z = Z_{\odot}$, the composite spectrum (red) of 726 luminous red galaxies at $0.47 < z < 0.55$ selected from the SDSS²¹ (available only for $\lambda > 2,600$ Å), the average spectra of $z \approx 1$ old (red) and dusty star-forming (green) EROs⁷, SSP synthetic spectra¹⁹ ($Z = Z_{\odot}$, Salpeter IMF) with ages of 0.5 Gyr (magenta), 1.1 Gyr (green) and 3.0 Gyr (red).

fitted without the need for dust extinction, and using a library of simple stellar population (SSP) models¹⁹ with a wide range of ages, metallicity $Z = Z_{\odot}$ and Salpeter initial mass function (IMF). This procedure yielded best-fitting ages of 1.0–1.7 Gyr, the mass-to-light ratios and hence the stellar mass of each galaxy, which results in the range of $(1-3) \times 10^{11} h_{70}^{-2} M_{\odot}$. $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$ (with $h_{70} = H_0/70$), $\Omega_m = 0.3$ and $\Omega_{\Lambda} = 0.7$ are adopted.

In addition to spectroscopy, the nature of these galaxies was investigated with the fundamental complement of Hubble Space Telescope + ACS (Advanced Camera for Surveys) imaging from the GOODS public Treasury Program¹⁷. The analysis of the ACS high-resolution images reveals that the surface brightness distribution of these galaxies is typical of elliptical/early-type galaxies (Fig. 4).

Besides pushing to $z \approx 1.9$ the identification of the highest-redshift elliptical galaxy, these objects are very relevant to the understanding of the evolution of galaxies in general for three main reasons: their old age, their high mass, and their substantial number density.

Indeed, an average age of about 1–2 Gyr ($Z = Z_{\odot}$) at $\langle z \rangle \approx 1.7$ implies that the onset of the star formation occurred not later than at $z \approx 2.5$ – 3.4 ($z \approx 2$ – 2.5 for $Z = 2.5 Z_{\odot}$). These are strict lower limits because they follow from assuming instantaneous bursts, whereas a more realistic, prolonged star-formation activity would push the bulk of their star formation to an earlier cosmic epoch. As an illustrative example, the photometric SED of ID 646 ($z = 1.903$)

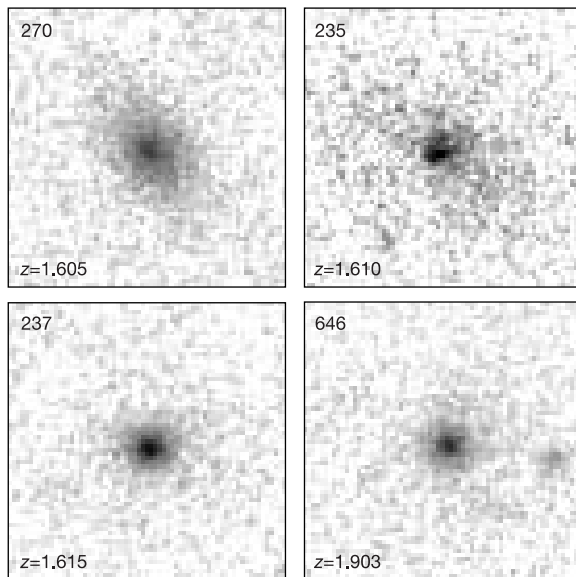


Figure 4 The morphological properties of the detected galaxies. Images of the four galaxies taken with the Hubble Space Telescope + ACS through the F850LP filter (from GOODS data¹⁷), which samples the rest-frame $\sim 3,000$ – $3,500 \text{ \AA}$ for $1.6 < z < 2$. The images are in logarithmic grey-scale and their size is $2'' \times 2''$, corresponding to $\sim 17 \times 17 \text{ kpc}$ for the average redshift $z = 1.7$ and the adopted cosmology. At a visual inspection, the galaxies show rather compact morphologies with most of the flux coming from the central regions. A fit of their surface brightness profiles was performed with a ‘Sersic law’ ($\propto r^{1/n}$) convolved with the average point spread function extracted from the stars in the ACS field and using the GASPHOT²⁹ and GALFIT³⁰ software packages. Objects ID 237 and ID 646 have profiles with acceptable values of n in the range of $4 < n < 6$, that is, typical of elliptical galaxies, object ID 270 is better reproduced by a flatter profile ($1 < n < 2$), whereas a more ambiguous result is found for the object showing some evidence of irregularities in the morphology (ID 235, $1 < n < 3$). These latter objects may be bulge-dominated spirals but no bulge/disk decomposition was attempted. Ground-based near-infrared images taken under $0.5''$ seeing conditions with the ESO VLT + ISAAC through the K_s filter (rest-frame $\sim 6,000$ – $8,000 \text{ \AA}$) show very compact morphologies, but no surface brightness fitting was done.

can be reproduced (without dust) with either a ~ 1 -Gyr-old instantaneous burst occurring at $z \approx 2.7$, or with a ~ 2 -Gyr-old stellar population with a star-formation rate declining with $\exp(-t/\tau)$ ($\tau = 0.3 \text{ Gyr}$). In the latter case, the star-formation onset would be pushed to $z \approx 4$ and half of the stars would be formed by $z \approx 3.6$. In addition, with stellar masses $M_{\star} > 10^{11} h_{70}^{-2} M_{\odot}$, these systems would rank among the most massive galaxies in the present-day Universe, suggesting that they were already fully assembled at this early epoch.

Finally, their number density is very high. Within the co-moving volume relative to 32 arcmin^2 and $1.5 < z < 1.9$ ($40,000 h_{70}^{-3} \text{ Mpc}^3$), the co-moving density of such galaxies is about $10^{-4} h_{70}^{-3} \text{ Mpc}^{-3}$, corresponding to a stellar mass density of about $2 \times 10^7 h_{70} M_{\odot} \text{ Mpc}^{-3}$, that is, about 10% of the local ($z = 0$) value²² for masses greater than $10^{11} M_{\odot}$. This mass density is comparable to that of star-forming $M_{\star} > 10^{11} M_{\odot}$ galaxies at $z \approx 2$ (ref. 23), suggesting that while the most massive galaxies in the local Universe are now old objects with no or weak star formation, by $z \approx 2$ passive and active star-forming massive galaxies coexist in nearly equal number.

Although more successful than previous models, the most recent realizations of semi-analytic hierarchical merging simulations still severely underpredict the density of such old galaxies: just one old galaxy with $K_s < 20$, $R - K_s > 6$, and $z > 1.5$ is present in the mock catalogue⁴ for the whole five times wider GOODS/CDFS area.

As expected for early-type galaxies^{9,24}, the three galaxies at $z \approx 1.61$ may trace the underlying large-scale structure. In this case, our estimated number density may be somewhat biased toward a high value. On the other hand, the number of such galaxies in our sample is likely to be a lower limit owing to the spectroscopic redshift incompleteness. There are indeed up to three more candidate old galaxies in the GOODS/K20 sample with $18.5 \leq K_s \leq 19.5$, $1.5 \leq z_{\text{phot}} \leq 2.0$, $5.6 \leq R - K_s \leq 6.8$ and compact HST morphology. Thus, in the GOODS/K20 sample the fraction of old galaxies among the whole $z > 1.5$ galaxy population is $15 \pm 8\%$ (spectroscopic redshifts only), or up to $25 \pm 11\%$ if also all the 3 additional candidates are counted.

It is generally thought that the so-called ‘redshift desert’ around $1.4 < z < 2.5$ represents the cosmic epoch when most star-formation activity and galaxy mass assembly took place²⁵. Our results show that, in addition to actively star-forming galaxies²⁶, a substantial number of ‘fossil’ systems also already populate this redshift range, and hence remain undetected in surveys biased towards star-forming systems. The luminous star-forming galaxies found at $z > 2$ in sub-millimetre²⁷ and near-infrared^{23,28} surveys may represent the progenitors of these old and massive systems. $\square$

Received 13 February; accepted 17 May 2004; doi:10.1038/nature02668.

1. Fukugita, M., Hogan, C. J. & Peebles, P. J. E. The cosmic baryon budget. *Astrophys. J.* **503**, 518–530 (1998).
2. Renzini, A. in *The Formation of Galactic Bulges* (eds Carollo, C. M., Ferguson, H. C. & Wyse, R. F. G.) 9–26 (Cambridge Univ. Press, Cambridge, UK, 1999).
3. Peebles, P. J. E. in *A New Era in Cosmology* (eds Metcalfe, N. & Shanks, T.) 351–361 (ASP Conf. Proc. Vol. 283, Astronomical Society of the Pacific, San Francisco, 2002).
4. Somerville, R. S. *et al.* The redshift distribution of near-infrared-selected galaxies in the Great Observatories Origins Deep Survey as a test of galaxy formation scenarios. *Astrophys. J.* **600**, L135–L139 (2004).
5. Freedman, W. L. & Turner, M. S. Colloquium: Measuring and understanding the universe. *Rev. Mod. Phys.* **75**, 1433–1447 (2003).
6. Im, M. *et al.* The DEEP Groth strip survey. X. Number density and luminosity function of field E/SO galaxies at $z < 1$. *Astrophys. J.* **571**, 136–171 (2002).
7. Cimatti, A. *et al.* The K20 survey. I. Disentangling old and dusty star-forming galaxies in the ERO population. *Astron. Astrophys.* **381**, L68–L73 (2002).
8. Benítez, N. *et al.* Detection of evolved high-redshift galaxies in deep NICMOS/VLT images. *Astrophys. J.* **515**, L65–L69 (1999).
9. Daddi, E. *et al.* Detection of strong clustering of extremely red objects: implications for the density of $z > 1$ ellipticals. *Astron. Astrophys.* **361**, 535–549 (2000).
10. Zepf, S. E. Formation of elliptical galaxies at moderate redshifts. *Nature* **390**, 377–380 (1997).
11. Rodighiero, G., Franceschini, A. & Fasano, G. Deep Hubble Space Telescope imaging surveys and the formation of spheroidal galaxies. *Mon. Not. R. Astron. Soc.* **324**, 491–497 (2001).
12. Taniguchi, Y. *et al.* Lyman α emitters beyond redshift 5: The dawn of galaxy formation. *J. Korean Astron. Soc.* **36**, 123–144 (2003).

13. Dunlop, J. S. *et al.* 3.5-Gyr-old galaxy at redshift 1.55. *Nature* **381**, 581–584 (1996).
14. Spinrad, H. *et al.* LBDS 53W091: an old, red galaxy at $z = 1.552$. *Astrophys. J.* **484**, 581–601 (1997).
15. Gavazzi, G., Pierini, D. & Boselli, A. The phenomenology of disk galaxies. *Astron. Astrophys.* **312**, 397–408 (1996).
16. Cimatti, A. *et al.* The K20 survey. III. Photometric and spectroscopic properties of the sample. *Astron. Astrophys.* **392**, 395–406 (2002).
17. Giavalisco, M. *et al.* The Great Observatories Origins Deep Survey: Initial results from optical and near-infrared imaging. *Astrophys. J.* **600**, L93–L98 (2004).
18. Pickles, A. J. A stellar spectral flux library: 1150–25000 Å. *Publ. Astron. Soc. Pacif.* **110**, 863–878 (1998).
19. Bruzual, G. & Charlot, S. Stellar population synthesis at the resolution of 2003. *Mon. Not. R. Astron. Soc.* **344**, 1000–1028 (2003).
20. Jimenez, R. *et al.* Synthetic stellar populations: single stellar populations, stellar interior models and primordial proto-galaxies. *Mon. Not. R. Astron. Soc.* **349**, 240–254 (2004).
21. Eisenstein, D. J. *et al.* Average spectra of massive galaxies in the Sloan Digital Sky Survey. *Astrophys. J.* **585**, 694–713 (2003).
22. Cole, S. *et al.* The 2dF galaxy redshift survey: near-infrared galaxy luminosity functions. *Mon. Not. R. Astron. Soc.* **326**, 255–273 (2001).
23. Daddi, E. *et al.* Near-infrared bright galaxies at $z \sim 2$. Entering the spheroid formation epoch? *Astrophys. J.* **600**, L127–L131 (2004).
24. Davis, M. & Geller, M. J. Galaxy correlations as a function of morphological type. *Astrophys. J.* **208**, 13–19 (1976).
25. Dickinson, M., Papovich, C., Ferguson, H. C. & Budavari, T. The evolution of the global stellar mass density at $0 < z < 3$. *Astrophys. J.* **587**, 25–40 (2003).
26. Steidel, C. C. *et al.* A survey of star-forming galaxies in the $z = 1.4$ – 2.5 ‘redshift desert’: Overview. *Astrophys. J.* **604**, 534–550 (2004).
27. Genzel, R. *et al.* Spatially resolved millimeter interferometry of SMM J02399–0136: A very massive galaxy at $z = 2.8$. *Astrophys. J.* **584**, 633–642 (2003).
28. Franx, M. *et al.* A significant population of red, near-infrared-selected high-redshift galaxies. *Astrophys. J.* **587**, L79–L83 (2003).
29. Pignatelli, E. & Fasano, G. GASPHOT: A tool for automated surface photometry of galaxies. *Astrophys. Space Sci.* **269**, 657–658 (1999).
30. Peng, C. Y., Ho, L. C., Impey, C. D. & Rix, H.-W. Detailed structural decomposition of galaxy images. *Astron. J.* **124**, 266–293 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work is based on observations made at the European Southern Observatory, Paranal, Chile, and with the NASA/ESA Hubble Space Telescope obtained at the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy (AURA). We thank R. Somerville for information on the GOODS/CDFS mock catalogue. We are grateful to the GOODS Team for obtaining and releasing the HST and FORS2 data

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.C. (cimatti@arcetri.astro.it).

An unusual isotope effect in a high-transition-temperature superconductor

G.-H. Gweon¹, T. Sasagawa^{2,3}, S.Y. Zhou⁴, J. Graf¹, H. Takagi^{2,3,5}, D.-H. Lee^{1,4} & A. Lanzara^{1,4}

¹Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

²Department of Advanced Materials Science, University of Tokyo, Kashiwa, Chiba 277-8561, Japan

³CREST, Japan Science and Technology Agency, Saitama 332-0012, Japan

⁴Department of Physics, University of California, Berkeley, California 94720, USA

⁵RIKEN (The Institute of Physical and Chemical Research), Wako 351-0198, Japan

In conventional superconductors, the electron pairing that allows superconductivity is caused by exchange of virtual phonons, which are quanta of lattice vibration. For high-transition-temperature (high- T_c) superconductors, it is far from clear that phonons are involved in the pairing at all. For example, the negligible change in T_c of optimally doped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ ($\text{Bi}2212$; ref. 1) upon oxygen isotope substitution ($^{16}\text{O} \rightarrow ^{18}\text{O}$)

leads to T_c decreasing from 92 to 91 K) has often been taken to mean that phonons play an insignificant role in this material. Here we provide a detailed comparison of the electron dynamics of $\text{Bi}2212$ samples containing different oxygen isotopes, using angle-resolved photoemission spectroscopy. Our data show definite and strong isotope effects. Surprisingly, the effects mainly appear in broad high-energy humps, commonly referred to as ‘incoherent peaks’. As a function of temperature and electron momentum, the magnitude of the isotope effect closely correlates with the superconducting gap—that is, the pair binding energy. We suggest that these results can be explained in a dynamic spin-Peierls picture², where the singlet pairing of electrons and the electron–lattice coupling mutually enhance each other.

We compare angle-resolved photoemission spectroscopy (ARPES) data of optimally doped $\text{Bi}2212$ samples at the three different stages of the isotope substitution loop $^{16}\text{O} \rightarrow ^{18}\text{O} \rightarrow ^{16}\text{O}$ (Supplementary Methods). In this way, we study directly the impact on the electron spectral function due to a modification of phonon properties, and thus gain insights into the nature of electron–phonon interaction in this material. Here we use the term ‘phonons’ loosely to denote quanta of lattice vibrations including spatially localized ones^{3,4}. To ensure that material properties unrelated to the isotope mass did not change during the isotope substitution process, we controlled the sample growth condition with high precision (Supplementary Methods) and checked the sample quality with various post-growth characterization tools, including ARPES itself (Supplementary Figure). All ARPES data were recorded at the Advanced Light Source as detailed elsewhere⁵.

Figure 1 shows low temperature (25 K) ARPES spectra and their dispersions along the nodal (ΓY) direction, where the superconducting gap is zero. In this and the rest of the figures, blue, red and green represent results for ^{16}O , ^{18}O and re-substituted $^{16}\text{O}_R$ ($^{18}\text{O} \rightarrow ^{16}\text{O}$) samples, respectively. In Fig. 1a, raw ARPES spectra as a function of energy (that is, the energy distribution curves, EDCs) are shown for different momenta along the nodal direction. Each EDC shows a peak, which sharpens up as momentum (k) approaches the normal state Fermi surface⁶. We loosely refer to a sharp peak as a ‘coherent peak’ (CP), and a broad hump as an ‘incoherent peak’ (IP)⁷. We have no intention of implying that the energy and the width of CPs satisfy the Landau quasiparticle requirement. Making a comparison between the blue and red curves in Fig. 1a, we detect definite but small isotope effect. The most notable change is the shift of the peak position in curves 4 and 5 by approximately 15 meV, which is bigger than our error bar by a factor of three. Inspecting curves 1–6 in Fig. 1a, we note that the isotope effect is maximum for binding energy in the range of 100–300 meV, and vanishes as energy decreases (cut 1) and increases (cut 6) from this region. We note that this energy range coincides with range of J to $2J$, where J is the super-exchange interaction of neighbouring spins.

Figure 1b shows the dispersions derived from the ARPES spectra at fixed energies as a function of momentum, known as the momentum distribution curves (MDCs). As reported earlier^{6,8–12}, we observe a ‘kink’, that is, a change of the slope in the dispersion, at energy ~ 70 meV. This feature has been used as evidence that a bosonic mode renormalizes the electron dynamics¹³. A comparison between the ^{16}O and ^{18}O dispersions in Fig. 1b clearly shows that the kink separates the low-energy regime where the spectra show CP and negligible isotope effect, from the high-energy regime where the spectra show IP and appreciable isotope effect. This observation suggests that phonons contribute appreciably to the electron self energy. To illustrate the size of our error bar, we show the dispersion of the $^{18}\text{O} \rightarrow ^{16}\text{O}$ re-substituted sample as the green line in Fig. 1b. Clearly the experimental uncertainties caused by isotope substitution are smaller than the isotope-induced changes reported here. We note that the insensitivity to isotope substitution at low energies is consistent with the notion of ‘universal nodal quasiparticle

properties' inferred from the doping independence of nodal Fermi velocity¹¹. The marked increase of the EDC width above the kink energy (Fig. 1a and ref. 6) and the concurrent strengthening of the isotope effect suggest that this broadening is at least partly due to scattering of electrons by phonons. Figure 1b inset shows the real part of the electron self energy (Σ) as a function of energy (ω), $\text{Re}\Sigma(\omega)$, obtained from measuring the deviation of the experimental dispersion from a common straight line representing the band structure dispersion¹⁴. We regard $\text{Re}\Sigma(\omega)$ mainly as a tool to zoom in on the isotope-induced changes in the dispersion. As expected, the overall isotope-induced change in $\text{Re}\Sigma(\omega)$ extends up to very high energy. Additionally, the maximum position of $\text{Re}\Sigma(\omega)$, that is, the kink energy¹⁴, shows an isotope-induced redshift.

Both our EDC and MDC analyses show that phonons have important effect on the electron dynamics in the energy range 100–300 meV. We note that in the same energy range it is believed that magnetic fluctuations also contribute importantly to the electron self energy¹⁵. We also note that the fact that the isotope effect of $\text{Re}\Sigma(\omega)$ extends up to very high energy goes well beyond the Migdal–Eliashberg model, where the phonon effect is confined near the kink energy¹⁶.

In Fig. 2a, we report the evolution of the isotope effect as the electron momentum approaches the antinodal region near the M point, where the *d*-wave superconducting gap reaches its maximum (see Fermi surface diagram in Fig. 2d inset). For cuts 2–6, where a

non-zero superconducting gap exists, the MDC dispersion is shown only up to the gap edge. We defer the discussion of the isotope effect on the gap to later in this Letter. At low energy, a kink in the dispersion can be clearly identified for all cuts (see arrows), becoming stronger for near-antinodal cuts^{5,12,17}. For all cuts, the kinks show an isotope induced redshift, $(5\text{--}10) \pm 5$ meV. As in Fig. 1, the kink energy defines a crossover from the low-energy regime where the spectra show CP and negligible isotope effect, to the high-energy regime where the spectra show IP and strong isotope effect. Again, this result suggests that phonons are indeed key players in causing the kink^{5,6,17}. At higher energy, the isotope-induced changes increase significantly as the momentum gets closer to the antinodal region. Moreover, a subtle sign change of the isotope effect is observed near cut 3. The results in Fig. 2a are confirmed by raw MDC data—for example, those shown in Fig. 2b. As noted earlier, isotope-induced changes are fully reversible upon isotope re-substitution (green line).

In Fig. 2c we show the raw EDC spectra at the momentum value

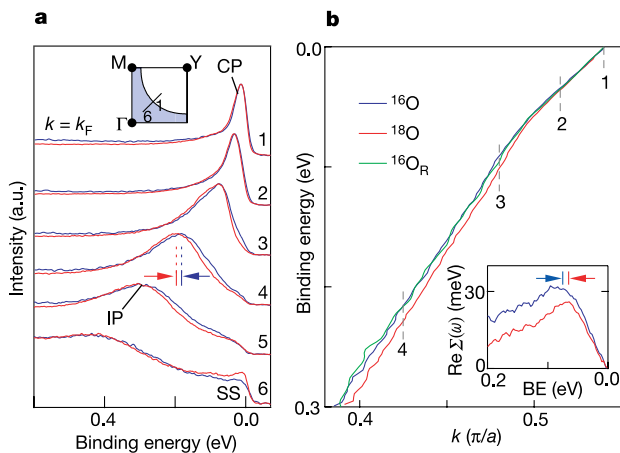


Figure 1 Isotope-induced changes of the nodal dispersion. The ARPES data were taken on optimally doped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ samples with different oxygen isotopes at $T = 25$ K, that is, in the superconducting state, along the nodal (Γ –Y) direction. **a**, Isotope dependence of raw EDCs. Inset shows a quadrant of the Brillouin zone, divided by the Fermi surface (curve) into electron-filled (shaded) and electron-empty parts. The momentum values corresponding to EDCs are marked in the inset and in **b** (dashed lines). The spectra show a coherent peak (CP) at low energy, almost isotope-independent, and a broad hump, that is, an incoherent peak (IP), strongly isotope dependent. The two mix in the crossover region (curve 3). Throughout this Letter, we use 'energy' and 'binding energy' interchangeably. All curves were scaled to the same peak height, and vertically displaced by different amounts for easy viewing. A small peak at the Fermi energy (0) in curve 6 is the well-known superstructure (SS) replica of the main band (also, see Supplementary Figure). **b**, Isotope dependence of MDC dispersion, obtained by Lorentzian fit of MDCs. The symbol k denotes the momentum value parallel to the Γ –Y direction. Consistent with **a**, the low energy dispersion is nearly isotope-independent, while the high energy dispersion is isotope-dependent. The effect is fully reversed by isotope re-substitution (green). The fits are shown only up to 300 meV, because they become less accurate at high energies. Inset shows the real part of the electron self energy, $\text{Re}\Sigma(\omega)$, obtained from the MDC dispersion¹⁴ by subtracting a line approximation for one electron band $\epsilon(k)$, connecting two points, one at E_F and the other at 300 meV binding energy, of the ^{18}O dispersion. The kink position, defined as the binding energy of the peak in $\text{Re}\Sigma(\omega)$, undergoes a redshift upon isotope substitution (see arrows).

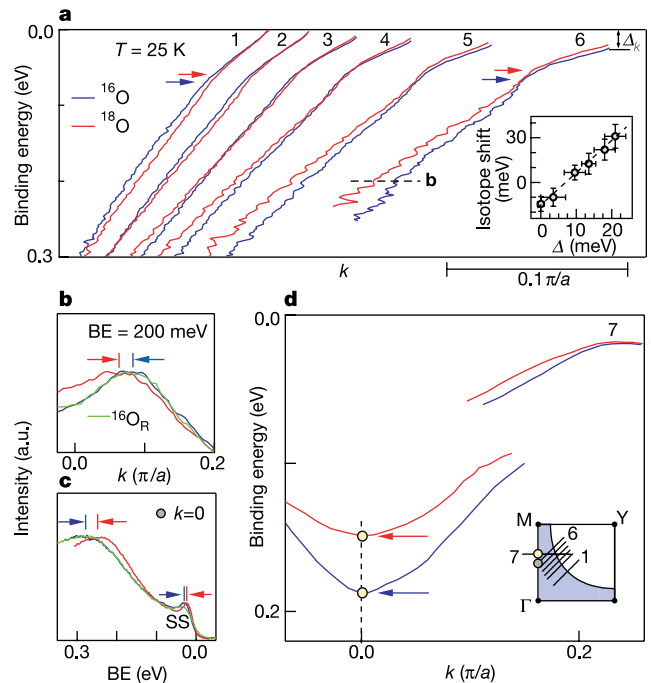


Figure 2 Isotope-induced changes of the off-nodal dispersions in the superconducting state. The ARPES data were taken on optimally doped $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$ samples with different oxygen isotopes at $T = 25$ K. In this figure, k denotes the momentum value parallel to the M–Y axis (inset in **d**). **a**, MDC dispersions for various cuts parallel to Γ –Y (see inset in **d**). A different origin of the momentum axis is used for each cut for an easy view of all data. Inset shows the isotope shift, measured at the momentum value for which the isotope-averaged binding energy is 220 meV, versus the superconducting gap (Δ), isotope-averaged. The strong linear correlation (dashed line is a guide to the eye) is independent of the binding energy used. **b**, Raw MDCs at binding energy (BE) = 200 meV for the data of cut 6, confirming the fit results of **a**. **c**, Raw EDCs at $k = 0$ (see inset in **d**) for cut 6. A large reversible isotope-induced shift (about -30 meV) of the EDC peak position is observed at high energy. Additionally, a small CP due to the well-known superstructure (SS) replica is observed (also, see Supplementary Figure). The large MDC and EDC shifts are fully reversible upon the isotope re-substitution process (green in panels **b** and **c**). **d**, EDC dispersions for cut 7 (inset), extracted from the maximum intensity positions of EDCs. The data show a large isotope-induced shift at high energy, similarly to the data in **a**. Note also that the EDC dispersion shows a qualitatively different behaviour from the MDC dispersion, for example, those shown in **a**, namely completely separate low-energy and high-energy branches. This difference stems from the fact that near the kink EDC line shape shows a two-peak structure (for example, Fig. 2 of ref. 6) while MDC line shape continues to show a single peak.

marked with a grey-filled circle in Fig. 2d inset. Here a shift of about -30 meV of the IP is observed. This value is a factor of 2 larger than the maximum shift reported along the nodal direction (see cut 1 and Fig. 1a). The weak low energy CP in each EDC of Fig. 2c is the well-known weak superstructure (SS) replica of the main band CP (also, see Supplementary Figure). Their isotope shift (about -6 meV) is an order of magnitude smaller than that of the IP at higher energy. Clearly, this strongly energy-dependent isotope effect cannot be ascribed to an overall spectral shift or be described by the Migdal–Eliashberg theory. The increase of the spectral shift with increasing binding energy suggests that the multi-phonon effect^{18,19} contributes to the electron self energy up to very high energy²⁰. Again, we note that in the high-energy regime not only the electron–electron interaction^{21,22} but also the electron–phonon interaction²³ affect the electron dynamics.

The inset of Fig. 2a summarizes the momentum dependence of high-energy isotope shift. The isotope-induced shift is plotted as a function of the superconducting gap. The linear correlation suggests a strongly anisotropic isotope effect, which increases as the electron momentum approaches the antinodal region.

In Fig. 2d, we show the dispersion for a cut parallel to the M–Y direction, determined from the EDC peak position. The location of this cut is shown as cut 7 in the inset. As in Fig. 2a, it is remarkable that, the higher the binding energy is, the stronger is the isotope effect. As the bottom of the dispersion is reached (yellow-filled circles), an isotope shift (about -40 meV) nearly 20% of the entire width of the dispersion is observed (see arrows). The increase of the shift is consistent with Fig. 2a inset.

An additional important point is the isotope effect on the superconducting gap, Δ_k (see, for example, cuts 3–6 in Fig. 2a). Unlike the sample-independent, reproducible and reversible isotope effect at energies much larger than the gap energy, the changes of the superconducting gap are small and random in both magnitude and sign. In particular, the maximum value of the gap varies by about ± 5 meV from one sample to another, regardless of the isotope mass. This suggests that the main cause of gap modification is disorder²⁴, not the change in isotope mass. This is very different from the behaviour of conventional s-wave superconductors²⁵.

In Fig. 3, we study the effect of isotope substitution above the superconducting transition temperature ($T = 100$ K). Figure 3a shows the MDC dispersions along cuts 1 and 3–6 of Fig. 2a. Comparing Fig. 3a with Fig. 2a, we note an overall decrease of the isotope-induced changes. For example, in the antinodal region (that is, cuts 5 and 6) the isotope effect is markedly different below and above T_c , consistent with the raw MDCs and EDCs in Fig. 3b, c. This interesting finding—that the strength of electron–lattice interaction is strongly temperature dependent—invalidates previous assumptions^{9,10,15} that a strong temperature dependence of spectroscopic features rules out phonons as explanation. As the last detailed point, we note that the sign reversal of the high-energy isotope effect near cut 3 persists at high temperature, despite the reduced overall isotope effect.

Before discussing the implications of our results, we address two possible sources of experimental error: unintentional doping change induced by the substitution process, and sample misalignment. Both of them are ruled out by the temperature dependence and by the reversibility and the reproducibility of the observed effects upon repeated measurements. Two more arguments can be used to rule out doping differences (also, see Supplementary Figure). First, a direct comparison with the MDC dispersion at different doping shows that the sign reversal at high energy cannot be induced by a doping change. Second, if the changes in the high-energy nodal dispersions in Fig. 1 are due to a doping change Δx , $\Delta x \approx 0.05$ (ref. 11) is implied, which is five times bigger than the maximum doping uncertainty ($\Delta x \approx 0.01$) of the isotope substitution process.

A potentially simple explanation of our findings is the effect of

static lattice distortion, induced by the isotope substitution. However, it is unlikely that the small change in the oxygen mass can affect the overall lattice structure, as shown by diffraction results for $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ (ref. 26). Moreover, the fact that the isotope effects occur with opposite signs for different cuts in momentum space excludes a net average lattice change as explanation. Finally, the isotope substitution might induce a random static structural disorder, connected with the disorder seen in scanning tunnelling microscopy²⁴. However, whereas the observed fluctuation of the superconducting gap is consistent with this possibility, the larger and fully reproducible isotope-induced changes at high energy are not.

On the basis of these findings we propose the following model for the nature of electron–lattice coupling in high- T_c superconductors. The fact that the isotope effect on the ARPES spectral function becomes much stronger below T_c suggests a picture where pairing of electrons enhances their coupling to the lattice and vice versa, as in spin-Peierls physics². In this picture, the motion of electron pairs modifies the lattice distortion locally. If the coupling between the electron pair and the lattice is too strong, the pair will be localized and the lattice distortion becomes static. In that limit, the system becomes an insulator rather than a superconductor. In the interesting situation where the coupling is not that strong, the dynamic spin-Peierls distortion follows the coherent motion of electron pairs in a superconducting state. Creation of a photo-hole in this state

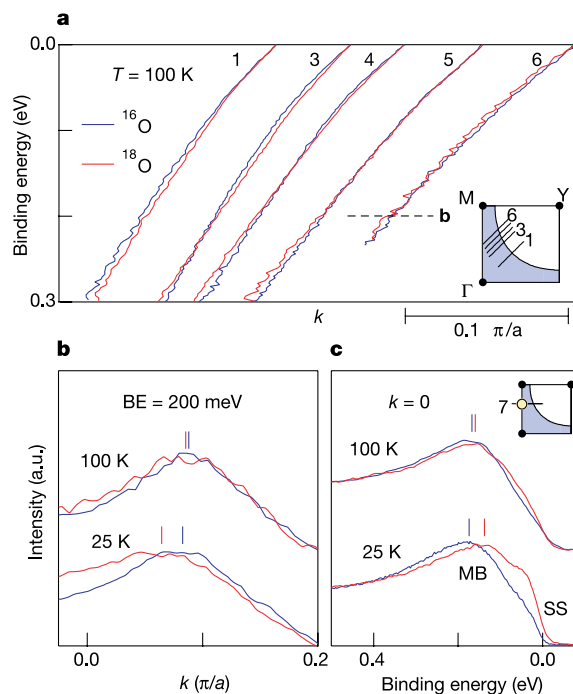


Figure 3 Decrease of the isotope-induced changes in the normal state. The samples, the k -axis definition, and the location of the cuts are identical to those in Fig. 2. **a**, Normal state MDC dispersions for cuts parallel to the Γ –Y direction, from Γ –Y to near the M point, measured at $T = 100$ K. A different origin of the momentum axis is used for each cut for easy viewing of all data. **b**, Comparison of the isotope-induced changes of the raw MDCs measured at 100 K and 25 K. The MDCs are measured at 200 meV binding energy for cut 6, as shown by the horizontal dashed line in **a**. **c**, Comparison of the isotope-induced changes of the raw EDCs, at 100 K and 25 K. The momentum value for the EDCs corresponds to the yellow-filled circle in the inset. As in Fig. 2c, a small peak near the Fermi energy (0) due to the well-known superstructure (SS) replica (also, see Supplementary Figure) coexists with the larger peak at high energy due to the main band (MB). An apparent isotope-induced change of the height of the SS peak at 25 K is merely due to the change of the position of the MB peak. All panels show that the isotope-induced changes are reduced greatly at high temperature, except for near the nodal direction.

causes the lattice to lose its distortion, and a strong coupling between the hole and the lattice results. Extrapolating this picture, we expect the diminishing of the spin-Peierls distortion above the pseudogap temperature T^* (note that $T^* = T_c$ for optimally doped Bi2212). As a result, the coupling between the photo-hole and the lattice is weakened. The observation of significant isotope dependences of T^* (refs 27), J and various low-temperature spin properties^{28,29} supports this scenario. We believe that the above cooperative interplay between electron pairing and electron–lattice interaction outlines the role that phonons play in high-temperature superconductivity. The above proposal has been shown theoretically to work in one dimension³⁰ and in a two-leg Hubbard ladder incorporating the electron–phonon interaction (A. Seidel, H. H. Lin and D.-H.L., manuscript in preparation). □

Received 1 April; accepted 7 June 2004; doi:10.1038/nature02731.

1. Franck, J. P. in *Physical Properties of High Temperature Superconductors IV* (ed. Ginsberg, D. M.) 189–293 (World Scientific, Singapore, 1994).
2. Pytte, E. Peierls instability in Heisenberg chains. *Phys. Rev. B* **10**, 4637–4642 (1974).
3. Müller, K. A. On the oxygen isotope effect and apex anharmonicity in high- T_c cuprates. *Z. Phys. B* **80**, 193–201 (1990).
4. Bianconi, A. *et al.* Determination of the local lattice distortions in the CuO_2 plane of $\text{La}_{1.85}\text{Sr}_{0.15}\text{CuO}_4$. *Phys. Rev. Lett.* **76**, 3412–3415 (1996).
5. Gweon, G.-H., Zhou, S. Y. & Lanzara, A. Strong influence of phonons on the electron dynamics of $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *J. Phys. Chem. Solids* (in the press).
6. Lanzara, A. *et al.* Evidence for ubiquitous strong electron-phonon coupling in high-temperature superconductors. *Nature* **412**, 510–514 (2001).
7. Allen, J. W., Gweon, G.-H., Claessen, R. & Matho, K. Fermi liquids and non-Fermi liquids—The view from photoemission. *J. Phys. Chem. Solids* **56**, 1849–1853 (1995).
8. Bogdanov, P. V. *et al.* Evidence for an energy scale for quasiparticle dispersion in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$. *Phys. Rev. Lett.* **85**, 2581–2584 (2000).
9. Kaminski, A. *et al.* Renormalization of spectral lineshape and dispersion below T_c in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *Phys. Rev. Lett.* **86**, 1070–1073 (2001).
10. Johnson, P. D. *et al.* Doping and temperature dependence of the mass enhancement observed in the cuprate $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *Phys. Rev. Lett.* **87**, 177007–177010 (2001).
11. Zhou, X. J. *et al.* Universal nodal Fermi velocity. *Nature* **423**, 398 (2003).
12. Sato, T. *et al.* Observation of band renormalization effects in hole-doped high- T_c superconductors. *Phys. Rev. Lett.* **91**, 157003 (2003).
13. Eliashberg, G. M. Interactions between electrons and lattice vibrations in a superconductor. *Zh. Eksp. Teor. Fiz.* **38**, 966–976 (1960); *Sov. Phys. JETP* **11**, 696–702 (1960).
14. Verga, S., Knigavko, A. & Marsiglio, F. Inversion of angle-resolved photoemission measurements in high- T_c cuprates. *Phys. Rev. B* **67**, 054503 (2003).
15. Hwang, J., Timusk, T. & Gu, G. D. High-transition-temperature superconductivity in the absence of the magnetic-resonance mode. *Nature* **427**, 714–717 (2004).
16. Rotenberg, E., Schaefer, J. & Kevan, S. D. Coupling between adsorbate vibrations and an electronic surface state. *Phys. Rev. Lett.* **84**, 2925–2928 (2000).
17. Cuk, T. *et al.* Coupling of the B_{1g} phonon to the anti-nodal electronic states of $\text{Bi}_2\text{Sr}_2\text{Ca}_{0.92}\text{Y}_{0.08}\text{Cu}_2\text{O}_{8+\delta}$. Preprint condmat/0403521 at (<http://xxx.lanl.gov>) (2004); *Phys. Rev. Lett.* (submitted).
18. Zhao, G. M., Hunt, M. B., Keller, H. & Müller, K. A. Evidence for polaronic supercarriers in the copper oxide superconductors $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$. *Nature* **385**, 236–240 (1997).
19. Crespi, V. H. & Cohen, M. L. Anharmonic phonons and the anomalous isotope effect in $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$. *Phys. Rev. B* **44**, 4712–4715 (1991).
20. Sawatzky, G. A. Testing Fermi liquid models. *Nature* **342**, 480–481 (1989).
21. Shen, Z.-X. & Schrieffer, J. R. Momentum, temperature, and doping dependence of photoemission lineshape and implications for the nature of the pairing potential in high- T_c superconducting materials. *Phys. Rev. Lett.* **78**, 1771–1774 (1997).
22. Norman, M. R. *et al.* Unusual dispersion and line shape of the superconducting state spectra of $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *Phys. Rev. Lett.* **79**, 3506–3509 (1997).
23. Paci, P. *et al.* Polaronic and nonadiabatic phase diagram from anomalous isotope effects. Preprint condmat/0405053 at (<http://xxx.lanl.gov>) (2004).
24. Pan, S. H. *et al.* Microscopic electronic inhomogeneity in the high- T_c superconductor $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_{8+\delta}$. *Nature* **413**, 282–285 (2001).
25. Anderson, P. W. Theory of dirty superconductors. *J. Phys. Chem. Solids* **11**, 26–30 (1959).
26. Crawford, M. K. *et al.* Oxygen isotope effect and structural phase transitions in La_2CuO_4 -based superconductors. *Science* **250**, 1390–1394 (1990).
27. Lanzara, A. *et al.* Oxygen-isotope shift of the charge-stripe ordering temperature in $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ from X-ray absorption spectroscopy. *J. Phys. Condens. Matter* **11**, L541–L546 (1999).
28. Shengelaya, A., Brunn, M., Kochelaev, B. I., Safina, A., Conder, K. M. & Müller, K. A. A metallic phase in lightly doped $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ observed by electron paramagnetic resonance. Preprint cond-mat/0310152 at (<http://xxx.lanl.gov>) (2003); *Phys. Rev. Lett.* (in the press).
29. Zech, D., Keller, H., Conder, K., Kaldis, E., Liarakis, N. & Müller, K. A. Site-selective oxygen isotope effect in optimally doped $\text{YBa}_2\text{Cu}_3\text{O}_{6+x}$. *Nature* **371**, 681–683 (1994).
30. Zimanyi, G. T., Kivelson, S. A. & Luther, A. Superconductivity from predominantly repulsive interactions in quasi one-dimensional systems. *Phys. Rev. Lett.* **60**, 2089–2092 (1988).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to K. A. Müller, A. Bianconi, N. L. Saini, D. Pines, A. Bill, V. Z. Kresin, S. A. Kivelson, A. J. Leggett, J. Clarke, J. Orenstein, M. L. Cohen, L. Pietronero, E. Cappelluti, J. C. Davis, J. W. Allen, A. S. Alexandrov, J. C. Phillips, A. H. Castro Neto, C. Castellani, A. Bussman Holder, D. Mihailovic, G. Deutscher, C. Bernhard, S. Uchida and T. Schneider for discussions. We thank Z. X. Shen, Z. Hussain, D. S. Chemla and N. V. Smith for support in the initial stage of the project. The work at UC Berkeley and LBNL was supported by the Department of Energy's Office of Basic Energy Science, Division of Materials Science.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.L. (alanzara@lbl.gov).

Colloidal nanocrystal heterostructures with linear and branched topology

Delia J. Milliron^{1,2}, Steven M. Hughes^{1,2}, Yi Cui^{1,2}, Liberato Manna^{1,2*}, Jingbo Li³, Lin-Wang Wang³ & A. Paul Alivisatos^{1,2}

¹Department of Chemistry, University of California, ²Materials Science Division, and ³Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California, 94720, USA

* Present address: National Nanotechnology Lab of INFN, Via Arnesano, 73100 Lecce Lecce, Italy

The development of colloidal quantum dots has led to practical applications of quantum confinement, such as in solution-processed solar cells¹, lasers² and as biological labels³. Further scientific and technological advances should be achievable if these colloidal quantum systems could be electronically coupled in a general way. For example, this was the case when it became possible to couple solid-state embedded quantum dots into quantum dot molecules^{4,5}. Similarly, the preparation of nanowires with linear alternating compositions—another form of coupled quantum dots—has led to the rapid development of single-nanowire light-emitting diodes⁶ and single-electron transistors⁷. Current strategies to connect colloidal quantum dots use organic coupling agents^{8,9}, which suffer from limited control over coupling parameters and over the geometry and complexity of assemblies. Here we demonstrate a general approach for fabricating inorganically coupled colloidal quantum dots and rods, connected epitaxially at branched and linear junctions within single nanocrystals. We achieve control over branching and composition throughout the growth of nanocrystal heterostructures to independently tune the properties of each component and the nature of their interactions. Distinct dots and rods are coupled through potential barriers of tuneable height and width, and arranged in three-dimensional space at well-defined angles and distances. Such control allows investigation of potential applications ranging from quantum information processing to artificial photosynthesis.

Unlike vapour–liquid–solid (VLS)- or SLS-grown nanowires, anisotropic nanocrystals in homogeneous solutions grow without the benefit of catalyst activation of one end. Hence, heterostructure growth in colloidal nanocrystals has so far been limited to core–shell structures that serve primarily to further isolate quantum dots from their environment^{10,11,12,13,14}. An elegant extension of core–shell growth enabled concentric alternating layers of CdS and HgS, which have a type I (nested) band alignment^{15,16}. Control over the electronic structure of concentric heterostructures is, however, restricted by their simple geometry and by strain due to lattice mismatch, which typically limits the thickness of each layer to a few monolayers or less. Heterostructures based on nanorods permit

more complexity and their properties are fully tuneable because strain does not limit their dimensions.

Anisotropic colloidal heterostructures are fabricated by sequential growth of semiconductor dots and rods of different materials, with the potential for branched connectivity in each generation. Branching is introduced through crystal phase control^{17,18,19}, so the large class of semiconductors exhibiting zincblende–wurtzite polytypism²⁰ could be incorporated into branched heterostructures by these methods. Recently, limited phase control enabled the high-yield synthesis of tetrapod-shaped nanocrystals of a single material, CdTe¹⁹, effectively arranging four quantum rods of the same composition around a central dot. This fundamental branched structure results from nucleation in the cubic zincblende phase with subsequent anisotropic growth in the hexagonal wurtzite phase. Here, we demonstrate that branched and linear junctions can be created not just at nucleation, but at any point during the growth of heterostructures. Considering two generations of growth within this paradigm, four basic structures can be postulated and were synthesized. The first-generation nanostructures can be linear, wurtzite rods, or branched tetrapods. On these two basic structures, a second material is grown in either branched or linear fashion (Figs 1 and 2). The dimensions of each generation define the degree of quantum confinement and are controlled by methods developed for nanorod growth²¹. The terminal rods and dots are coupled through the tuneable barrier defined by the first generation, whereas more generations of growth would produce structures incorporating even more complex interactions. In a preliminary exploration

of novel properties of nanocrystal heterostructures, long-range photoinduced charge separation has been achieved in type II (staggered band) heterostructures, as shown by the quenching of nanocrystal luminescence. Type I heterostructures permit tuneable exchange coupling between the terminal quantum dots or rods.

New methods were developed to grow a second material selectively on the ends of nanorods and to create branch points at will. Unlike core–shell nanocrystals, heterostructures were grown in the kinetic control regime previously exploited for CdSe nanocrystal shape control¹⁷. First-generation structures were grown by modifications of previously reported methods for preparing elongated nanocrystals^{17,18,19,21}. In all cases, cadmium oxide complexed by alkylphosphonic acids was used as the cation precursor¹⁸. The anion precursor (elemental Se, S or Te dissolved in tri(*n*-alkylphosphine)) was injected at a temperature between 280 and 320 °C to initiate growth. Linear heterojunctions were formed when precursors for a second material were added to a growth solution containing preformed nanorods or tetrapods. Branched junctions were introduced preferentially at high supersaturation of these precursors. The second generation was typically grown without isolation of the nanocrystals by using an excess of cadmium in the first step and injecting the anion precursor for the second material. Thus, in this work, all the reported heterostructures have cadmium as a common cation. Linearly extended rods were synthesized with CdS rods and CdSe extensions, while the other structures were synthesized with CdSe in the first and CdTe in the second generation. Branched rods were also synthesized with CdS rods and CdTe branches. Similar

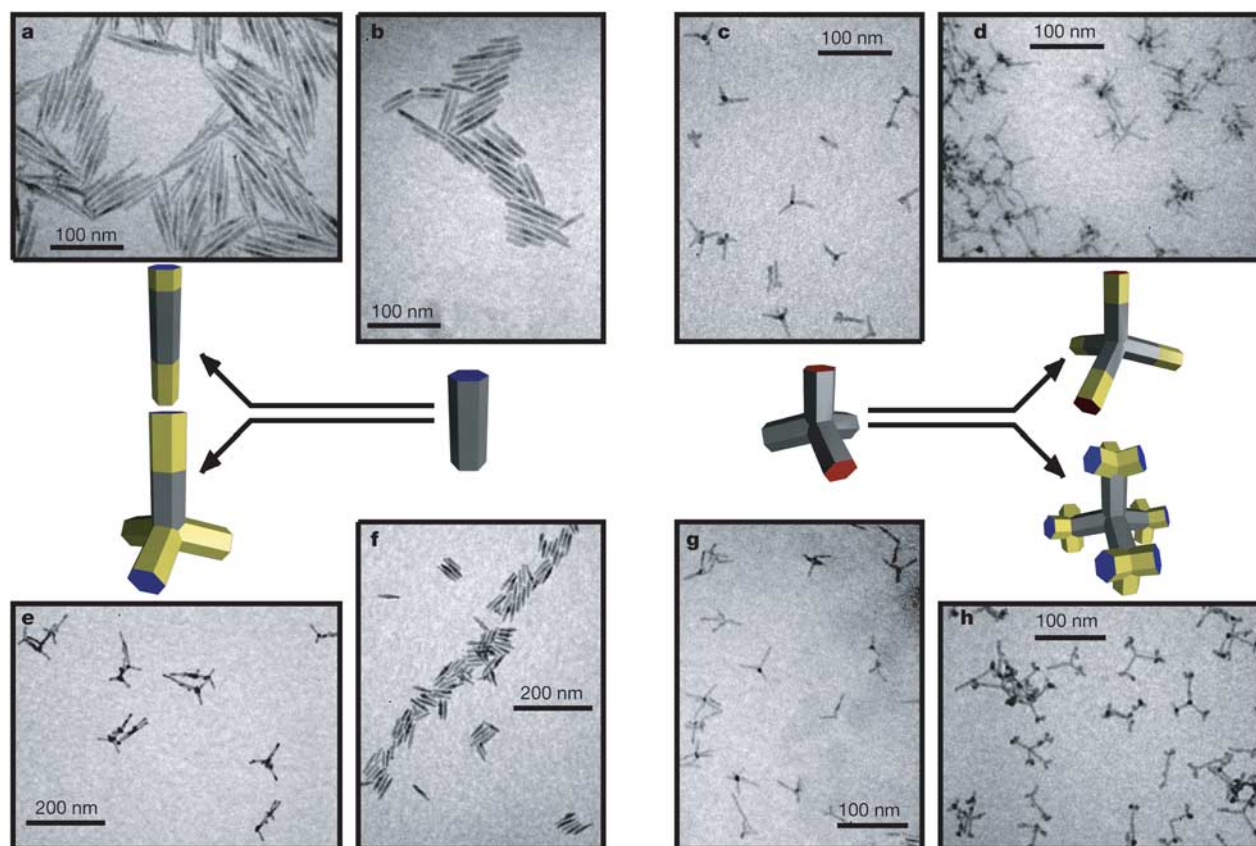


Figure 1 Survey of nanocrystal heterostructures. Extended rods (**a**) were formed first by growing CdS nanorods (**b**), then adding CdSe extensions to each end. Branched rods (**e**) result from nucleation of CdTe on either end of CdSe nanorods (**f**). A CdTe zincblende region at one end creates the branch point. CdSe tetrapods (**c**, **g**) comprise a zincblende core and four wurtzite arms. Extending each arm linearly with CdTe yields extended

tetrapods (**d**). Branched tetrapods (**h**) result from nucleation of CdTe zincblende branch points on the end of each arm. The model sequences illustrate the growth processes. In each case, the first generation is shown in grey and the second in yellow, while red and blue indicate the unique crystal faces of the wurtzite structure.

results were achieved by isolation of the first-generation nanocrystals and reintroduction of cadmium precursor before injecting the second anion precursor, implying extensibility to heterostructures incorporating semiconductors that do not share common ions. This method may also produce more compositionally abrupt heterojunctions. All of the heterostructures could be readily dispersed in common organic solvents, such as toluene and chloroform, and were prepared with high yield without any post-synthetic separation.

Reversing the growth sequence under similar synthetic conditions changed the growth pattern of the second material from selective on the ends to more homogeneous core-shell growth. For example, growing CdTe followed by CdSe, we succeeded in synthesizing type II core-shell tetrapods (Supplementary Fig. 1). As observed previously for core-shell nanorods²², elongated growth from the ends of core-shell tetrapods proceeds only after several monolayers of shell have formed and strain limits further homogeneous growth. In contrast, in branched and linear heterostructures, we suggest that the difference in surface energy between two materials favours end-selective growth. Although, in principle, they limit materials selection for a given topology, these observations suggest that any pair of materials in this class²⁰ could be combined in core-shell or in end-selective structures by reversing their growth order.

Several techniques were applied to establish the end selectivity of heteroepitaxy in these nanocrystal structures. Nanoprobe X-ray energy dispersive spectrometry (EDS) was used to examine the local atomic composition of the heterostructures³. The resulting EDS line scans (Fig. 3) confirm the presence of Te at either end of branched-rod heterostructures, Se in the central rod, and Cd throughout. The apparent remnant tellurium signal in the middle section is an artefact that results from partial overlap with an adjacent Cd peak and is observed in similar magnitude in CdSe nanorods containing no Te. Results on extended rods (Supplementary Fig. 2) similarly confirm end long growth of thinner CdSe extensions on CdS rods. While a sufficiently small spot size was used, the spatial resolution of the EDS data was limited by the drift of the instrument so that it remains uncertain how compositionally abrupt the interfaces are. The apparently few-nanometre width of

the composition gradient is an upper bound. To see more clearly whether the end long growth is accompanied by the formation of a thin shell, the powder X-ray diffraction patterns at different stages of growth were examined (Supplementary Fig. 3). Although noticeable peak shifts due to strain result from the growth of uniform

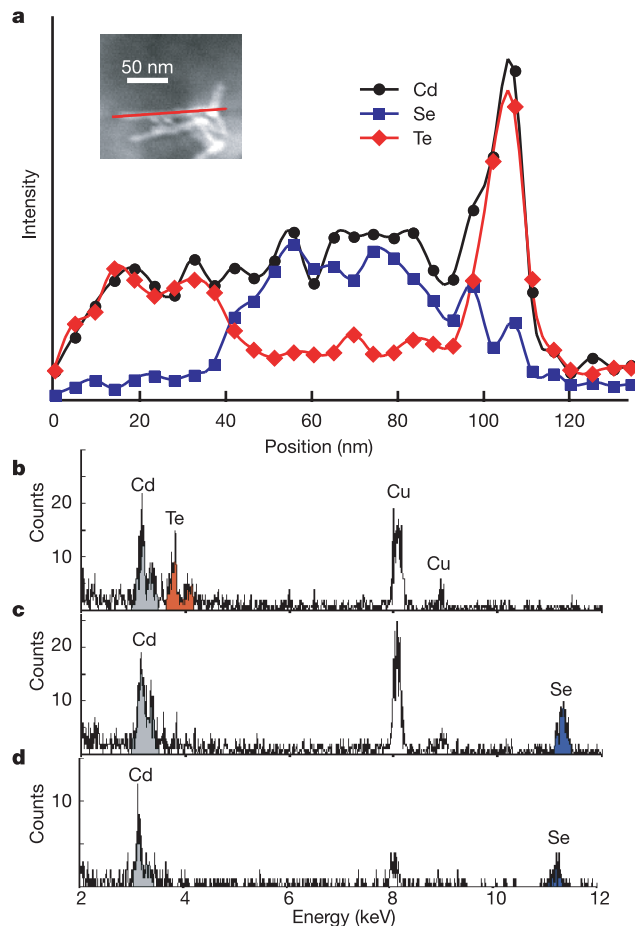


Figure 3 Analytical electron microscopy of heterojunctions. **a**, A nanoprobe EDS line scan (**a**) along a branched rod (inset) confirms CdTe growth on CdSe. The high-intensity spike corresponds to the CdTe arm projecting out of the plane from the branch point. **b–d**, Representative spectra from the CdTe portion (**b**) and CdSe portion (**c**) of the heterostructure, and shown together with a spectrum of a pure CdSe rod (**d**).

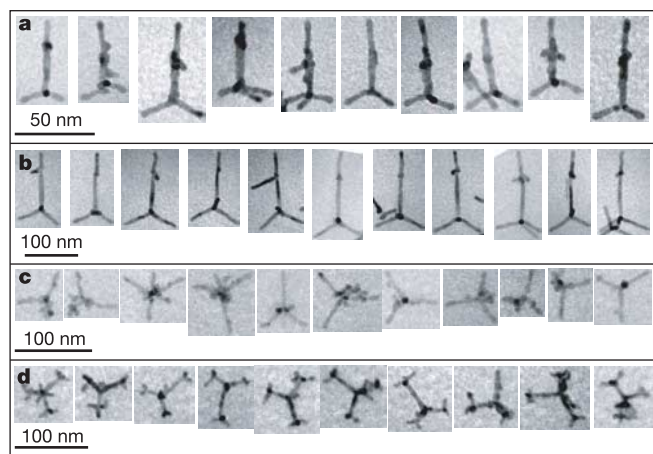


Figure 2 A closer look at nanocrystal heterostructures. Isolated particles allow examination of structural aspects such as ‘back branching’ in branched rods (**a** and **b**) and extended tetrapods (**c**) and structural isomers in branched tetrapods (**d**). Branches projecting from the linear junction at an angle back along the original rod arise from stacking faults at the heterojunction and are consistent with the symmetry relationships between the zincblende and wurtzite phases. Secondary branches in branched tetrapods (**d**) grow either staggered or eclipsed with respect to the arms of the original tetrapod. Heterostructures in **a**, **c** and **d** are CdTe grown on CdSe, and in **b** are CdTe grown on CdS.

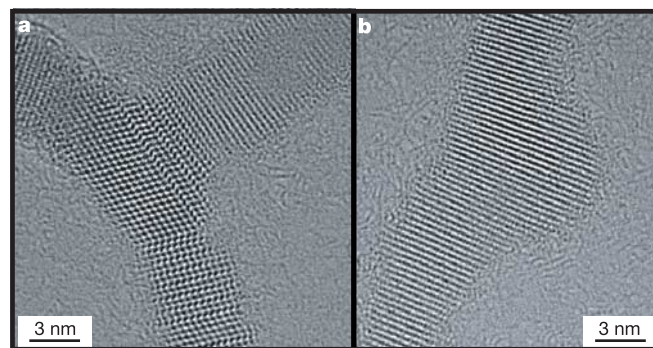


Figure 4 High-resolution electron microscopy of heterojunctions. **a**, HRTEM of a branch point shows a zincblende core and wurtzite branches of CdTe and an original wurtzite rod (upper right) of CdSe. **b**, Examination of a linear junction between CdSe (upper right) and CdTe reveals continuous wurtzite growth.

shells even one monolayer thick on nanorods^{22,23}, we observed no obvious peak shifts in the case where growth occurs on the ends only. Furthermore, statistical length and diameter distributions extracted from transmission electron microscope (TEM) images (Supplementary Fig. 4) indicate no significant change in diameter upon growth of the second material. For example, the radii of the arms of one batch of CdSe tetrapods were 2.9 ± 0.4 nm before growing CdTe extensions, and 2.9 ± 0.5 nm afterwards. CdS nanorod radii were 3.6 ± 0.4 nm before and 3.9 ± 0.5 nm after growing CdSe extensions where, owing to the tapered shapes of the CdS rods and of the heterostructures, the radius is taken as the maximum. The small increase in radius is consistent with the

formation of at most one monolayer of CdSe on the sides of the CdS nanorods. Finally, although several of these heterostructures are type II, their optical absorption spectra lack distinctive sub-band gap tails like those observed for type II core-shell quantum dots¹⁴ and for our core-shell tetrapods (Supplementary Fig. 5). Owing to the very small scale of these heterostructures, no technique allows us to eliminate the possibility that a very thin (one monolayer) shell grows by overgrowth or ion exchange; however, taken together, these results strongly indicate selective heteroepitaxial growth on the ends of nanorods and tetrapods.

The topology of each generation is determined by the initial growth phase of the nucleating material. Nanorods and tetrapod

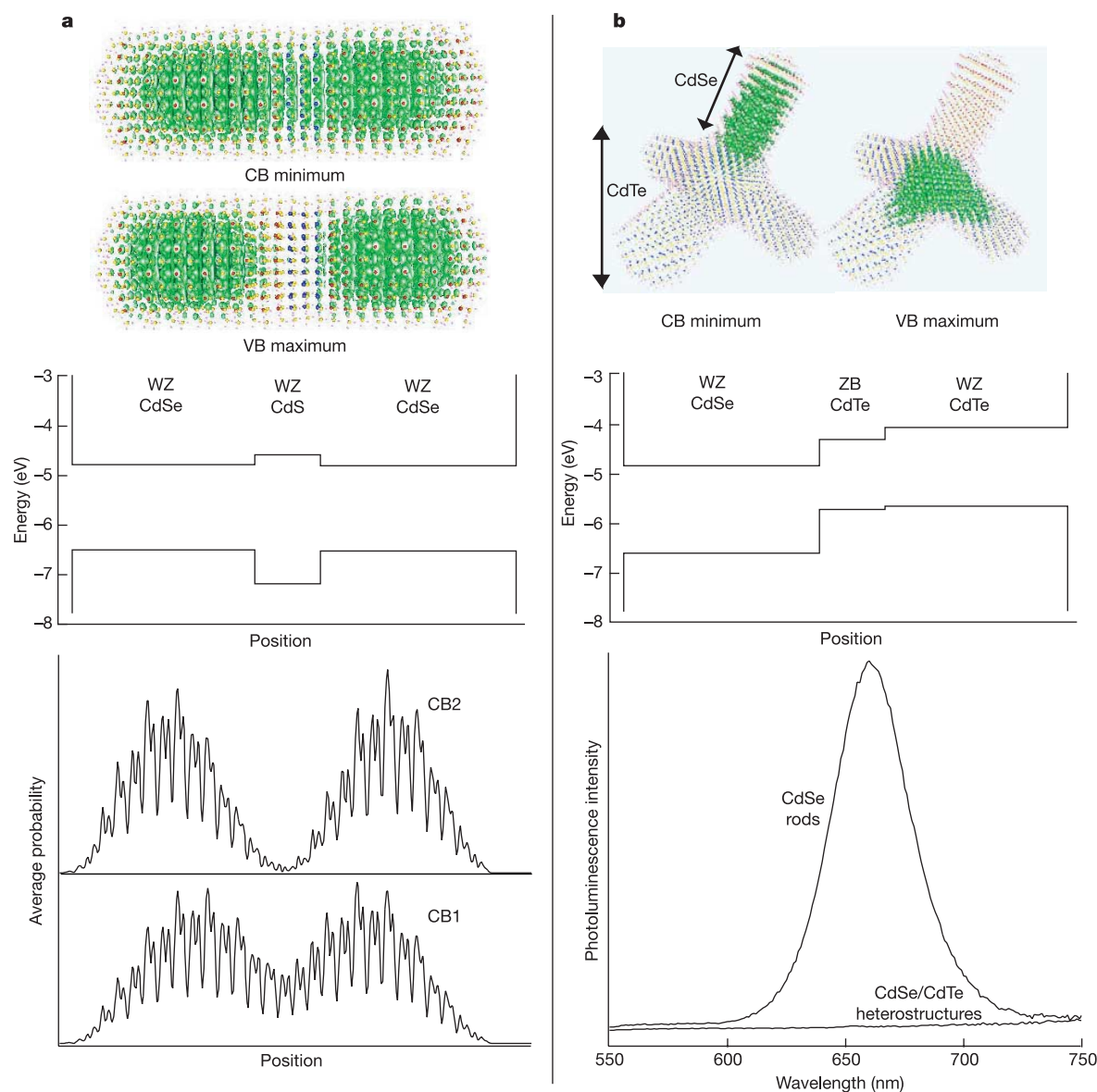


Figure 5 Optoelectronic properties of type I and type II heterostructures. **a**, *Ab initio* calculation of type I CdSe/CdS/CdSe heterostructures reveals electronic coupling. Isosurfaces of the lowest-energy electron and highest-energy hole states (top) show the even distribution of probability between the two terminal CdSe quantum rods. The band alignment illustrates the formation of coupled wells for electrons and holes. The cross section-averaged probability for the lowest-energy conduction band state (CB1) shows significant penetration of the CdS barrier, while the next conduction band state (CB2) has a longitudinal node. The diameter of the calculated rod is about 2 nm, the total length is

about 9.5 nm, and there are in total 2,002 atoms. The thickness of the CdS is three monolayers. **b**, In type II CdSe/CdTe heterostructures the electron and hole are spatially separated. Their distribution agrees qualitatively with the expected band alignment. The photoluminescence of CdSe rods immediately before adding CdTe branches is easily observed (bottom), whereas in heterostructures, charge separation strongly quenches the luminescence. For the calculated tetrapod, each branch has a diameter of 2.2 nm and a length of 4.2 nm. There are in total 3,685 atoms in the tetrapod.

arms grow in the wurtzite structure, elongated along the unique *c* axis. Invariably, such nanocrystals contain a statistical distribution of stacking faults, which convert the growth to zincblende and back to wurtzite along the rod, sometimes leading to kinks or other irregularities²¹. High-resolution transmission electron microscopy (HRTEM) of linear junctions found in extended rods and tetrapods, and in branched rods, reveals a continuation of anisotropic wurtzite growth in the second semiconductor (Fig. 4 and Supplementary Fig. 6), often accompanied by a high concentration of stacking faults. At these junctions, the small diameter allows dislocation-free epitaxial growth despite fairly large lattice mismatches. The heterostructures with the largest mismatch (CdS/CdTe) easily accommodated an 11% mismatch (Supplementary Fig. 6). In HRTEM, branch points could most easily be observed by imaging branched rods that were missing one of the three CdTe branches (Fig. 4). In such nanocrystals, a well-formed CdTe zincblende region could be seen at the junction. Thus, a branched junction forms when the new material initially grows in the zincblende structure, followed by a reversion to anisotropic wurtzite growth, forming the branches. Zincblende formation is favoured by a high supersaturation of the precursors immediately following injection, with wurtzite growth resuming as concentrations drop.

Our approach to synthesizing nanocrystal heterostructures can not only create solution processible analogues of nanowire heterostructures, but also enables unique functionality through the three-dimensional arrangement of components. Representative heterostructures reported here incorporate either type I or type II interfaces to define the nature of the interactions between components. In the first case, CdSe extensions grown on a CdS nanorod are quantum rods separated by a barrier for electrons and holes (Fig. 5a). The coupling of these rods is tuneable by changing the length of the original rod or of the extensions, or selecting a different material to vary the barrier height. *Ab initio* calculation of the electronic structure of model heterostructures confirms that the lowest energy level is split into 'symmetric' and 'anti-symmetric' combinations. A coupling energy of 27 meV was estimated for a heterostructure with a three-monolayer CdS barrier (three layers each of Cd and S) and this coupling energy drops to 9 meV when the CdS barrier is six monolayers thick. Such tuneable coupling, combined with the long spin coherence times observed in colloidal nanocrystals²⁴, make these heterostructures intriguing candidates for control of quantum coherence.

Branched tetrapods with CdSe central tetrapods and terminal CdTe branches are interesting for their unusual charge-separating properties. These structures absorb light across the visible spectrum, then separate electrons and holes across their type II interfaces. The sharp reduction in spatial overlap between electrons and holes, apparent in the *ab initio* result, effectively quenches the bandgap photoluminescence (Fig. 5b). Both CdSe rods and CdTe tetrapods emit well-defined bandgap luminescence under similar conditions. Calculations have furthermore suggested that electrons localize in the zincblende core of CdSe tetrapods²⁵. In the heterostructured branched tetrapods (Fig. 2d), this implies long-range charge separation with the electron at the CdSe zincblende core and the hole delocalized in the outer CdTe branches, 30 nm or more away. This distance can be tuned by the changing dimensions of the central tetrapod. Organic dendrimers designed for such radial charge separation²⁶ required three generations of growth and purification to separate charges by only a few nanometres. The nature and lifetime of this proposed charge-separated state is currently under investigation, as are possible applications to photovoltaic energy conversion. □

Methods

As an example of heterostructure synthesis, a typical preparation for CdSe/CdTe branched rods dissolved 104 mg of CdO in 0.81 g of octadecylphosphonic acid (PolyCarbon Industries) and 3.19 g of tri(*n*-octyl)phosphine (TOP) oxide at 300 °C under air-free

conditions. At 280 °C, 15.8 mg of selenium dissolved in 320 mg of TOP was injected and the CdSe nanorods grew for four minutes. Then, at 290 °C, 34 mg of tellurium dissolved in 306 mg of TOP was injected and CdTe branches grew for six minutes before the heat was removed to stop the reaction. The resulting heterostructures were deposited from solution onto carbon-coated copper TEM grids for analysis. Low-resolution TEM images were collected on a Philips Tecnai 12 microscope, while HRTEM employed a Philips CM300FEG/UT and a Philips Tecnai G2 20. EDS spectra and line scans were collected on a Philips CM200FEG STEM with a 0.5-nm spot size. Despite this spot size, the spatial resolution was limited by the drift of the instrument so that it remains unclear how compositionally abrupt the heterojunctions are.

The theoretical calculations were done using the charge patching method²⁷. This method calculates the band edge states of a nanosystem with *ab initio* accuracy, but scales linearly to the size of the system. The structures were idealized to incorporate abrupt heterojunctions because the actual composition variation is, at this point, unknown. However, we expect the theoretical results to hold qualitatively for less perfect heterostructures. The surface of a nanosystem is passivated with pseudo-hydrogen atoms, for example, one pseudo-hydrogen atom with nuclear charge $Z = 1.5$ electron for each Cd dangling bond, and $Z = 0.5$ electron for each Se, S, or Te dangling bond. This simple model represents an ideal passivation that captures the essence of any good experimental passivations. The atomic positions of a given passivated binary semiconductor nanosystem (that is, CdSe/CdTe, CdSe/CdS) are relaxed using the valence force field (VFF) method. This atomistic method describes the elastic aspects of the system accurately. After the atomic positions of a given system are determined, small prototype systems are precalculated under the local density approximation (LDA) of the density functional theory. The charge densities of these small prototype systems are used to generate localized charge motifs around each atom. The charge motifs for bulk Cd, Se, Te and S atoms and surface pseudo-hydrogen atoms, and the derivatives of these charge motifs regarding the change of bond lengths and bond angles are all generated. These charge motifs and their derivatives are then placed together to generate the LDA charge density of a given nanosystem. The details of this method are reported in ref. 27. The typical error of the so generated charge density compared to the directly calculated charge density under LDA is about 1%. The resulting eigenstate error is about 20–40 meV. The energy splittings between states within the conduction or valence band have typical errors of a few meV. After the charge density is obtained, LDA formulae are used to calculate the LDA total potential and the LDA hamiltonian. Then the folded spectrum method^{28,29} is used to calculate the band edge states. Under this procedure, the *ab initio* accuracy band edge states of a thousand-atom nanosystem can be calculated within a few hours on a parallel computer. To calculate coupling energies, a small external electrical field was applied to cancel the permanent dipole of the nanorod³⁰. The diameter of the calculated rod shown in Fig. 5a is about 2 nm, the total length is about 9.5 nm, and there are in total 2,002 atoms. For the calculated tetrapod in Fig. 5b, each branch has a diameter of 2.2 nm and a length of 4.2 nm. There are in total 3,685 atoms in the tetrapod.

Received 14 November 2003; accepted 28 May 2004; doi:10.1038/nature02695.

- Huynh, W. U., Dittmer, J. J. & Alivisatos, A. P. Hybrid nanorod-polymer solar cells. *Science* **295**, 2425–2427 (2002).
- Kazes, M., Lewis, D. Y., Ebenstein, Y., Mokari, T. & Banin, U. Lasing from semiconductor quantum rods in a cylindrical microcavity. *Adv. Mater.* **14**, 317–321 (2002).
- Bruchez, M., Moronne, M., Gin, P., Weiss, S. & Alivisatos, A. P. Semiconductor nanocrystals as fluorescent biological labels. *Science* **281**, 2013–2016 (1998).
- Schedelbeck, G., Wegscheider, W., Bichler, M. & Abstreiter, G. Coupled quantum dots fabricated by cleaved edge overgrowth: From artificial atoms to molecules. *Science* **278**, 1792–1795 (1997).
- Bayer, M. et al. Coupling and entangling of quantum states in quantum dot molecules. *Science* **291**, 451–453 (2001).
- Gudiksen, M. S., Lauhon, L. J., Wang, J., Smith, D. C. & Lieber, C. M. Growth of nanowire superlattice structures for nanoscale photonics and electronics. *Nature* **415**, 617–620 (2002).
- Thelander, C. et al. Single-electron transistors in heterostructure nanowires. *Appl. Phys. Lett.* **83**, 2052–2054 (2003).
- Collier, C. P., Vossmeier, T. & Heath, J. R. Nanocrystal superlattices. *Annu. Rev. Phys. Chem.* **49**, 371–404 (1998).
- Ouyang, M. & Awschalom, D. D. Coherent spin transfer between molecularly bridged quantum dots. *Science* **301**, 1074–1078 (2003).
- Hines, M. A. & Guyot-Sionnest, P. Synthesis and characterization of strongly luminescing ZnS-capped CdSe nanocrystals. *J. Phys. Chem.* **100**, 468–471 (1996).
- Dabbousi, B. O. et al. (CdSe)/ZnS core-shell quantum dots: Synthesis and characterization of a size series of highly luminescent nanocrystallites. *J. Phys. Chem. B* **101**, 9462–9475 (1997).
- Peng, X., Schlamp, M. C., Kadavanich, A. V. & Alivisatos, A. P. Epitaxial growth of highly luminescent CdSe/CdS core/shell nanocrystals with photostability and electronic accessibility. *J. Am. Chem. Soc.* **119**, 7019–7029 (1997).
- Cao, Y. & Banin, U. Growth and properties of semiconductor core/shell nanocrystals with InAs cores. *J. Am. Chem. Soc.* **122**, 9692–9702 (2000).
- Kim, S., Fisher, B., Eisler, H.-J. & Bawendi, M. Type II quantum dots: CdTe/CdSe(core/shell) and CdSe/ZnTe(core/shell) heterostructures. *J. Am. Chem. Soc.* **125**, 11466–11467 (2003).
- Eychmüller, A., Mews, A. & Weller, H. A quantum dot quantum well: CdS/HgS/CdS. *Chem. Phys. Lett.* **208**, 59–62 (1993).
- Borchert, H. et al. Photoemission study of onion like quantum dot quantum well and double quantum well nanocrystals of CdS and HgS. *J. Phys. Chem. B* **107**, 7486–7491 (2003).
- Manna, L., Scher, E. C. & Alivisatos, A. P. Synthesis of soluble and processable rod-, arrow-, teardrop-, and tetrapod-shaped CdSe nanocrystals. *J. Am. Chem. Soc.* **122**, 12700–12706 (2000).
- Peng, Z. A. & Peng, X. Nearly monodisperse and shape-controlled CdSe nanocrystals via alternative routes: Nucleation and growth. *J. Am. Chem. Soc.* **124**, 3343–3353 (2002).
- Manna, L., Milliron, D. J., Meisel, A., Scher, E. C. & Alivisatos, A. P. Controlled growth of tetrapod-branched inorganic nanocrystals. *Nature Mater.* **2**, 382–385 (2003).

20. Yeh, C. Y., Lu, Z. W., Froyen, S. & Zunger, A. Zincblende–wurtzite polytypism in semiconductors. *Phys. Rev. B* **46**, 10086–10097 (1992).
21. Peng, X. *et al.* Shape control of CdSe nanocrystals. *Nature* **404**, 59–61 (2000).
22. Manna, L., Scher, E. C., Li, L. S. & Alivisatos, A. P. Epitaxial growth and photochemical annealing of graded CdS/ZnS shells on colloidal CdSe nanorods. *J. Am. Chem. Soc.* **124**, 7136–7145 (2002).
23. Mokari, T. & Banin, U. Synthesis and properties of CdSe/ZnS core/shell nanorods. *Chem. Mater.* **15**, 3955–3960 (2003).
24. Gupta, J. A., Awschalom, D. D., Peng, X. & Alivisatos, A. P. Spin coherence in semiconductor quantum dots. *Phys. Rev. B* **59**, 10421–10424 (1999).
25. Li, J. & Wang, L. W. Shape effects of electronic states of nanocrystals. *Nano Lett.* **3**, 1357–1363 (2003).
26. Capistosti, G. J., Cramer, S. J., Rajesh, C. S. & Modarelli, D. A. Photoinduced electron transfer within porphyrin-containing poly(amide) dendrimers. *Org. Lett.* **3**, 1645–1648 (2001).
27. Wang, L. W. Charge-density patching method for unconventional semiconductor binary systems. *Phys. Rev. Lett.* **88**, 256402 (2002).
28. Wang, L. W. & Zunger, A. Solving Schrödinger's equation around a desired energy: Application to silicon quantum dots. *J. Chem. Phys.* **100**, 2394–2397 (1994).
29. Canning, A., Wang, L. W., Williamson, A. & Zunger, A. Parallel empirical pseudopotential electronic structure calculations for million atom systems. *J. Comp. Phys.* **160**, 29–41 (2000).
30. Li, L. S. & Alivisatos, A. P. Origin and scaling of the permanent dipole moment in CdSe nanorods. *Phys. Rev. Lett.* **90**, 097402 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements This work was supported by the US Department of Energy. Some high-resolution and analytical electron microscopy was performed at the National Center for Electron Microscopy (NCEM) with the help of E. C. Nelson and some low-resolution electron microscopy was performed at the Electron Microscope Laboratory at UCB with the help of M. Casula. For the theoretical calculations we used the National Energy Research Scientific Computing Center.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.P.A. (alivis@uclink.berkeley.edu).

Export of dissolved organic carbon from peatlands under elevated carbon dioxide levels

C. Freeman¹, N. Fenner¹, N. J. Ostle², H. Kang^{1*}, D. J. Dowrick¹, B. Reynolds³, M. A. Lock¹, D. Sleep², S. Hughes³ & J. Hudson⁴

¹School of Biological Sciences, University of Wales, Bangor LL57 2UW, UK

²Centre for Ecology and Hydrology, Lancaster Environment Centre, Lancaster LA1 4AB, UK

³Centre for Ecology and Hydrology, Bangor LL57 2UP, UK

⁴Centre for Ecology and Hydrology, Wallingford OX10 8BB, UK

* Present address: Department of Environmental Science and Engineering, Ewha University, Seoul 120-750, Korea

Peatlands represent a vast store of global carbon¹. Observations of rapidly rising dissolved organic carbon concentrations in rivers draining peatlands have created concerns that those stores are beginning to destabilize^{2,3}. Three main factors have been put forward as potential causal mechanisms, but it appears that two alternatives—warming^{2,4} and increased river discharge³—cannot offer satisfactory explanations⁵. Here we show that the third proposed mechanism, namely shifting trends in the proportion of annual rainfall arriving in summer⁶, is similarly unable to account for the trend. Instead we infer that a previously unrecognized mechanism—carbon dioxide mediated stimulation of primary productivity—is responsible. Under elevated carbon dioxide levels, the proportion of dissolved organic carbon derived from recently assimilated carbon dioxide was ten times higher than that of the control cases. Concentrations of dissolved organic carbon appear far more sensitive to environmental drivers that affect net primary productivity than those affecting decomposition alone.

Peat-accumulating wetlands have long been recognized as unbalanced ecosystems where the rate of photosynthetic production of organic matter exceeds that of its decomposition¹. These wetlands have accumulated a vast pool of organic carbon, and currently hold about 390–455 Pg (1 Pg = 10¹⁵ g) of terrestrial C or approaching one-third of the global soil carbon stock^{1,7}. Concerns have been expressed that these stores may destabilize, releasing their carbon in either gaseous^{1,8} or aqueous form^{2,5}. Those concerns appear to be supported by observations of rapidly rising dissolved organic carbon (DOC) concentrations in aquatic ecosystems draining peatland catchments². The trend appears to have been in place for at least 39 years (ref. 5), and is not confined to Europe. In North America, for example, DOC concentrations in streams flowing into a series of boreal lakes were found to increase by 30–80% during one 20-year study⁹.

The causes of the observed rises in riverine DOC concentrations are attracting considerable debate^{3,5,6}. Some have suggested that the trend is due to increased decomposition in response to rising temperatures^{2,5}. Yet field and laboratory studies indicate that DOC concentrations are not sufficiently responsive to warming^{2,4} to account for a 65% increase in DOC when temperatures have risen by far less than 1 °C. Experimental manipulations suggest that more than 10 °C of additional warming would be required to yield such a large rise in DOC^{2,4}—a level of change not anticipated by even the most pessimistic of IPCC scenarios. Moreover, long-term (21-year) field studies in Canada have failed to find positive correlations between DOC and temperature¹⁰. Discharge, another alternative³, seems equally unable to offer a universal explanation. Whereas some studies note that DOC export increases in association with rising discharge¹¹, other workers have found clear evidence of long-term rises in DOC concentrations without any indication of a trend in discharge^{5,6}. A modified hydrological explanation has since been proposed: this suggests that rather than discharge changes alone, it

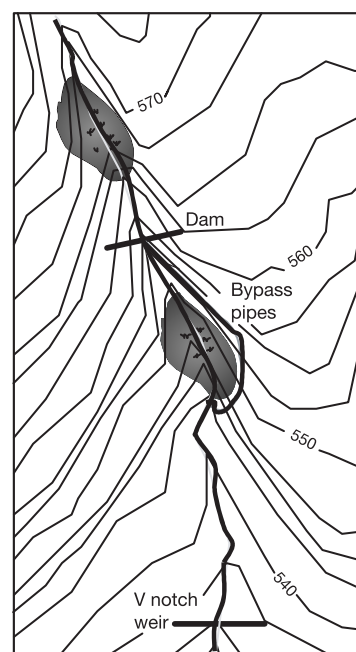


Figure 1 The field-based drought simulation. The hydrological effects of a reduction of summer rainfall were simulated by diverting waterflow around the lower of two wetlands (both wetlands shown shaded). The upper wetland was retained as a control. Contours shown as metres above sea level. The areas of the control and experimental wetlands were 348 m² and 336 m², respectively.

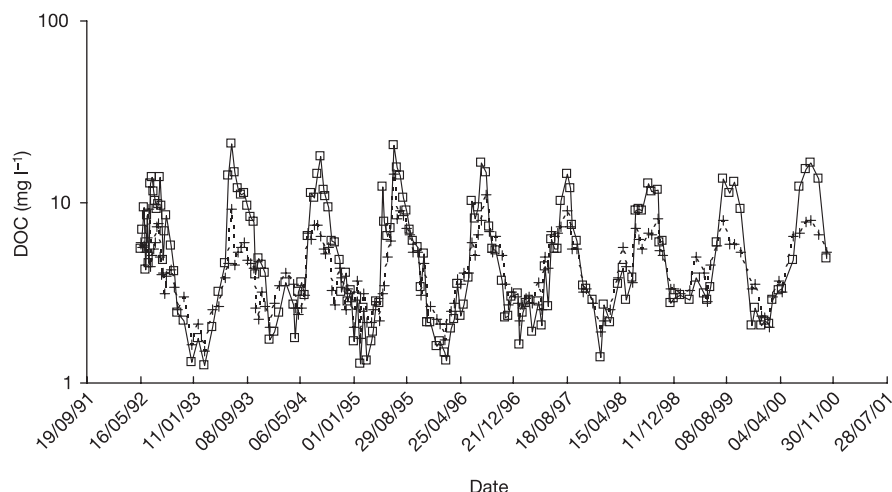


Figure 2 Comparison of DOC mobilization from control and drought-treated wetlands. Despite five simulations of the hydrological conditions associated with intense summer

drought, DOC release from the drought-affected system (dashed line) shows no evidence of sustained rises above those of the control system (solid line).

is a reduction in the proportion of annual precipitation arising in summer that might explain the long-term trend in DOC release¹². The hypothesis is based on observations that decomposition is accelerated under warmer, drier conditions and the assumption that soluble carbon compounds would be mobilized following the return to wetter conditions^{5,12}. But there is also a further environmental factor that has so far escaped consideration, and one that has shown consistent increases over the period of these studies, namely atmospheric CO₂ (ref. 13). The last two mechanisms have received insufficient consideration to be eliminated as drivers of the current DOC trend, and so formed the basis of our investigations.

We have compared the impacts of both drought-associated hydrological changes and elevated atmospheric CO₂ on DOC release from peatlands. To examine the impact of drought-associated hydrological changes, we simulated repeated summer droughts in a chain of peatlands in mid-Wales. A flow manipulation system was established (Fig. 1) with a control wetland, below which a dam was constructed to allow diversion of waterflows around an 'experimental' wetland to simulate summer droughts. Over a three-year period (1992–94) and then a further two-year period (1999–2000) the wetland was subjected to simulated droughts of varying duration (16–20 weeks) between late spring and early autumn. For the hypothesized role of hydrological conditions associated with reduced summer rainfall to remain tenable, a divergence in the levels of DOC released from each wetland would be anticipated, with the drought-affected wetland showing higher DOC release. However, each drought revealed peat DOC concentrations substantially lower than the control (averaging 44%, $P < 0.01$, in the early droughts (1992–94), for example). Such declines are to be expected as under the aerobic conditions associated with drought, gaseous CO₂, rather than dissolved organics, tends to be the major end product of decomposition. And despite five simulations of the drought-like hydrological changes, we found no evidence of DOC release increasing above that of the control, either during the manipulation or in the remainder of the year following each manipulation (Fig. 2). The hypothesized role of reduced summer precipitation in generating rising DOC trends therefore appears untenable.

It seemed more unlikely that the remaining alternative, elevated CO₂, would influence the mobilization of stored carbon from a peatland, at least not without indirect impacts such as warming² or modified rainfall patterns³. And yet in a comparison of the

responses of three Welsh peatlands to ambient and elevated (+235 p.p.m.) CO₂ (Fig. 3), we have found significantly greater DOC release from bog (14%, $P = 0.05$), fen (49%, $P < 0.05$) and riparian peatland (61%, $P < 0.01$) soils. The response is not a short-term transient impact¹⁴; measurements from the last site show the trend for rising DOC to persist for at least 3 years of exposure (Fig. 3, inset). Moreover, the ranked DOC releases (bog < fen < riparian) match rankings based on nutritional grounds¹⁵. Such responses cannot easily be explained solely in terms of mobilization of existing carbon stocks; decomposition is affected greatly by nutrient availability, but does not respond directly to elevated CO₂ (ref. 16).

On the basis of studies of terrestrial systems¹⁷, we propose that the nutrient dependence of this response to CO₂ indicates that the DOC increases were induced by increased primary production and DOC exudation from plants. When compared with relatively nutrient-rich wetlands (such as fens and riparian peatlands), bogs (as nutrient-poor ecosystems) are more likely to be constrained by nitrogen, phosphorus or certain other micronutrients¹⁸ than CO₂.

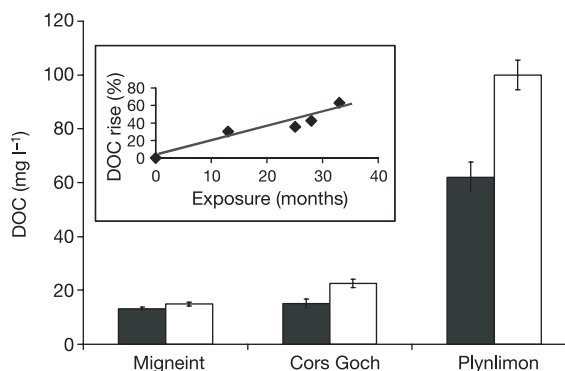


Figure 3 Effects of elevated CO₂ on DOC release. Under elevated CO₂, DOC release from three peatland sites of differing nutrient availability (Migneint bog, Cors Goch fen and Plynllyon riparian peatland) all increased significantly. The greatest increases were found in the sites with the highest nutrient status. Inset, monitoring temporal changes in DOC release from the Plynllyon samples revealed no evidence to suggest that the increased DOC release was a transient response. Error bars indicate standard errors.

As such, and as illustrated in Fig. 3, they are unlikely to show more than a modest response to CO₂ enrichment¹⁹. However, when nutrient availability increases in proportion to demand, net primary productivity (NPP) becomes less constrained and more responsive to elevated CO₂ (ref. 20), and thus more likely to leach additional carbon into the soil and associated river systems.

We tested our hypothesis that NPP was driving the increased DOC release by using ¹³C as a tracer. Intact peat monoliths that had been exposed to elevated CO₂ for nearly 3 years were pulse labelled with ¹³CO₂ (99 at.% ¹³C) for 5 hours to enable ¹³C enrichment by photosynthetic assimilation of CO₂-C. We then followed the translocation of pulse-derived C through the plant and into leachate DOC. ¹³C assimilation into the vegetation amounted to 22.9 and 35.8 mg under ambient and elevated CO₂, respectively. Under elevated CO₂, while the exudation of ¹³C-DOC only reached 0.6% of the plant-assimilated carbon, the proportion of DOC in the soil solution derived from recently assimilated CO₂ was ten times higher than that of the control (Fig. 4). The magnitude of this change suggests that DOC release is far more sensitive to atmospheric CO₂ than to warming² or the hydrological changes associated with droughts⁴ (Fig. 2).

The primary response of wetland plants to enhanced CO₂ may be increased root exudation, rather than an increase in photosynthesis or an increase in above-ground plant biomass. This proposal is supported by observations that while the response of photosynthesis and biomass to elevated CO₂ in wetlands is very mixed, most studies show an increase in methane fluxes^{21–26}, a process closely tied to below-ground plant carbon release²⁷. Other studies have found that elevated CO₂ modifies the abundance of vascular plants relative to mosses in peatlands²³. This was confirmed in our own study, where the cover of rooted vascular plants increased from 52% to 67% of the vegetation, and this too may have contributed to the increased DOC release. Recent work has shown that DOC release from peatlands can account for >15% of photosynthetic carbon capture²⁸. DOC release is further complicated by a 'priming mechanism', by which labile carbon released by roots stimulates microbial activity, leading to enhanced degradation of soil organic matter²⁹. A critical question remains, however; why should peatlands stand apart from other ecosystems in showing such marked rises in DOC export—surely all ecosystems should show this response? To answer this, we must remember that mobilized peatland DOC enters a medium that is largely anaerobic. First, under such conditions, released carbon is less likely to be completely metabolized to CO₂, and instead forms substantial amounts of dissolved organic end products. Second, these conditions represent a potent constraint for

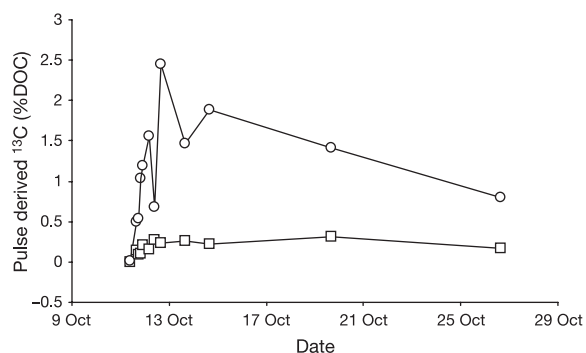


Figure 4 ¹³C pulse labelling of released DOC. ¹³C tracers confirmed that the elevated CO₂ treatment (circles) markedly increased the amount of DOC available for export owing to the release of recently photosynthesized DOC into the soil. Labelling of the control cores (ambient CO₂) is indicated by squares. The data were collected in 2001.

oxygen-dependent enzymes such as phenol oxidase⁸, one of the few enzymes able to degrade highly recalcitrant phenolic compounds such as lignins or humic and fulvic acids that form major decomposition products. Phenolic compounds accumulate and, acting as inhibitors^{8,30}, they can suppress enzymic hydrolysis and microbial metabolism both in the peatland and its recipient stream waters³⁰. The net effect is that carbon is more likely to be released in dissolved organic form from peatlands than from other ecosystems.

Rising DOC could present some benefits for aquatic ecosystems; increased DOC concentrations would create additional substrates for microbial food chains³⁰, while attenuating UV light to protect organisms from harmful wavelengths⁹ (although, in some places, the trend for rising DOC could be cancelled out by increased drought frequencies or acidification⁹). However, from a human perspective it is arguably water supplies that are of greatest concern; the presence of DOC results in water of low aesthetic quality, increases the threat of biological contamination of treated waters by consuming free residual chlorine, and can cause the formation of potentially carcinogenic trihalomethanes⁵. Such issues are becoming an increasing cause for concern in the water industry. □

Methods

Drought simulation experiment

The catchment chosen for study, Cerrig-yr-Wyn in the Upper Wye catchment on Plynlimon (UK National Grid Ref. SN 820 866), is a small gully, typical of many in the uplands of Wales in which flush wetlands have developed as discontinuous serial cascade systems. The gully mire (peat pH range 4.0–4.8) is characterized by *Sphagnum* and *Juncus* communities. Water movement consists mainly of flow in a meandering channel, alternately subterranean or exposed, located at the interface between the peat and the underlying bedrock (consisting of base-poor Ordovician mudstones or shales). The individual patches of wetland are well defined, making it relatively simple to control the supply of stream water to individual areas of wetland in order to artificially manipulate the hydrological conditions in the gully mire.

The water table within each patch of flush wetland is controlled naturally by a combination of recharge and discharge of the peat in the vicinity of the soil pipes and the surface balance between rainfall inputs, vertical and lateral drainage and evapotranspiration. A flow manipulation system was set up with a control wetland at the head of the system fed by rainfall and streamflow. Below the control wetland, a dam was constructed to allow diversion of stream flow through 150-mm pipes around an 'experimental' wetland to simulate summer droughts. The artificial barrier was not fully watertight, and this allowed some seepage to reach the experimental wetland during base flow conditions to allow realistic rates of compensation flow and to prevent complete drying. The rewetting of the experimental wetland after each drought simulation was achieved by distributing surface recharge using streamflow through 150-mm pipes across the head of the wetland.

Peat-water samples were extracted, at a minimum of monthly intervals, at 10-cm depth in the peat profile using soil solution samplers placed at 2.5-m spacing along a transect down the two wetlands. Five samplers were installed in each of the two wetlands. The approximate sampling time for each extracted sample was 2–3 min. The soil solution samplers were constructed using a design aimed at minimizing the extraction volume and any unnecessary dead volumes. This ensured that excessive disturbance of the wetland was avoided and that the integrity of the extracted sample (in particular the redox potential) was unaffected. These were constructed by removing the ends of ten 2.5-cm³ Plastipak syringes 1 cm from the luer tip. The luer end was retained and packed with glass wool and connected by autoanalyser transmission tubing to the surface, where suction could be applied using a 30-cm³ Plastipak syringe, into which the samples were collected.

On return to the laboratory, samples were filtered through 0.2 µm diameter membrane syringe filters and DOC analysed by Skalar autoanalyser following a UV-digestion technique and colorimetric detection. The validity of the analytical data generated by the laboratory was monitored by participation in a regular inter-laboratory proficiency scheme (WRC Aquacheck).

Elevated CO₂ experiments

Sample collection and exposure to elevated CO₂. Intact peat monoliths (11 cm diameter × 20 cm deep) were collected from three sites: (1) Migneint, a bog near Ffestiniog in north Wales that is nutrient-poor and receives nutrients solely from rainfall; (2) Cors Goch, a fen located on the island of Anglesey, north Wales, which gains more nutrients as a consequence of influxes from surrounding soils and ground water; and (3) a riparian peatland in the Upper Wye catchment of Plynlimon, mid-Wales. Riparian wetlands gain even better access to nutrients by intercepting nutrient-laden waters passing from terrestrial systems into recipient drainage streams.

The peat monoliths were exposed to ambient or elevated (+235 p.p.m.) CO₂ in the NERC-CEH Bangor Solardome facility. DOC in the soil solutions was determined using a total organic carbon analyser (Shimadzu 5000).

¹³C labelling

Peat plant-soil monoliths were subjected to ¹³C labelling in October 2001, having been in

the Solardomes since December 1998. For each treatment level, two natural abundance cores (NAB) and three labelled (^{13}C) cores were used. Vegetation was ^{13}C labelled by feeding 99.9 at.% ^{13}C - CO_2 into transparent acrylic labelling chambers. Leachate samples were taken at regular intervals before, during, and after ^{13}C labelling.

Stable isotope $^{13}\text{C}/^{12}\text{C}$ analyses

Analytical determinations of $^{13}\text{C}/^{12}\text{C}$ ratios were made at the NERC Stable Isotope Facility. Plant matter was analysed using a Roboprep Elemental Analyser coupled with a Tracermass IRMS. Headspace gas obtained from the oxidation of leachate and standard solution subsamples was also analysed by gas chromatograph (Autosystem XL) immediately after ^{13}C analysis in order to measure CO_2 concentrations and quantify recovery. Ratios of $^{13}\text{C}/^{12}\text{C}$ in headspace CO_2 samples were determined using a trace gas pre-concentration unit coupled to a Micromass Isoprime IRMS.

Received 18 September 2003; accepted 2 June 2004; doi:10.1038/nature02707.

- Gorham, E. Northern peatlands: role in the carbon cycle and probable responses to climatic warming. *Ecol. Appl.* **1**, 182–195 (1991).
- Freeman, C., Evans, C. D., Monteith, D. T., Reynolds, B. & Fenner, N. Export of organic carbon from peat soils. *Nature* **412**, 785 (2001).
- Tranvik, L. J. & Jansson, M. Climate change — Terrestrial export of organic carbon. *Nature* **415**, 861–862 (2002).
- Pastor, J. *et al.* Global warming and the export of dissolved organic carbon from boreal peatlands. *Oikos* **100**, 380–386 (2003).
- Worrall, F., Burt, T. & Shedden, R. Long term records of riverine dissolved organic matter. *Biogeochemistry* **64**, 165–178 (2003).
- Evans, C. D., Freeman, C., Monteith, D. T., Reynolds, B. & Fenner, N. Climate change—Terrestrial export of organic carbon—Reply. *Nature* **415**, 862 (2002).
- Jenkinson, D. S., Adams, D. E. & Wild, A. Model estimates of CO_2 emissions from soil in response to global warming. *Nature* **351**, 304–306 (1991).
- Freeman, C., Ostle, N. & Kang, H. An enzymic latch on a global carbon store. *Nature* **409**, 149 (2001).
- Schindler, D. W. *et al.* Climate-induced changes in the dissolved organic carbon budgets of boreal lakes. *Biogeochemistry* **36**, 9–28 (1997).
- Hudson, J. J., Dillon, P. J. & Somers, K. M. Long-term patterns in dissolved organic carbon in boreal lakes: the role of incident radiation, precipitation, air temperature, southern oscillation and acid deposition. *Hydrol. Earth Syst. Sci.* **7**, 390–398 (2003).
- Forsberg, C. Will an increased greenhouse impact in Fennoscandia give rise to more humic and coloured lakes? *Hydrobiologia* **229**, 51–58 (1992).
- Tipping, E. *et al.* Climatic influences on the leaching of dissolved organic matter from upland UK moorland soils, investigated by a field manipulation experiment. *Environ. Int.* **25**, 83–95 (1999).
- Houghton, J. T. *et al.* (eds) *Climate Change 2001: The Scientific Basis* (Cambridge Univ. Press, Cambridge, 2001).
- Oechel, W. C. *et al.* Transient nature of CO_2 fertilization in arctic tundra. *Nature* **371**, 500–503 (1994).
- Mitsch, W. J. & Gosselink, J. G. *Wetlands* (Van Nostrand Reinhold, New York, 1993).
- Norby, R. J., Cotrufo, M. F., Ineson, P., O'Neill, E. G. & Canadell, J. G. Elevated CO_2 , litter chemistry, and decomposition: a synthesis. *Oecologia* **127**, 153–165 (2001).
- Jones, T. H. *et al.* Impacts of rising atmospheric carbon dioxide on model terrestrial ecosystems. *Science* **280**, 441–443 (1998).
- Aerts, R., Wallen, B. & Malmer, N. Growth-limiting nutrients in *Sphagnum*-dominated bogs subject to low and high atmospheric nitrogen supply. *J. Ecol.* **80**, 131–140 (1992).
- Woodin, S., Graham, B., Killick, A., Skiba, U. & Cresser, M. Nutrient limitation of the long-term response of heather [*Calluna-vulgaris* (L) hull] to CO_2 enrichment. *New Phytol.* **122**, 635–642 (1992).
- Zangerl, A. R. & Bazzaz, F. A. The response of plants to elevated CO_2 . 2. Competitive interactions among annual plants under varying light and nutrients. *Oecologia* **62**, 412–417 (1984).
- Hutchin, P. R., Press, M. C., Lee, J. A. & Ashenden, T. W. Elevated concentrations of CO_2 may double methane emissions from mires. *Glob. Change Biol.* **1**, 125–128 (1995).
- Van der Heijden, E., Jauhiainen, J., Silvola, J., Vasander, H. & Kuiper, P. J. C. Effects of elevated atmospheric CO_2 concentration and increased nitrogen deposition on growth and chemical composition of ombrotrophic *Sphagnum balticum* and oligo-mesotrophic *Sphagnum papillosum*. *J. Bryol.* **22**, 175–182 (2000).
- Berendse, F. *et al.* Raised atmospheric CO_2 levels and increased N deposition cause shifts in plant species composition and production in *Sphagnum* bogs. *Glob. Change Biol.* **7**, 591–598 (2001).
- Dacey, J. W. H., Drake, B. G. & Klug, M. J. Stimulation of methane emission by carbon dioxide enrichment of marsh vegetation. *Nature* **370**, 47–49 (1994).
- Megonigal, J. P. & Schlesinger, W. H. Enhanced CH_4 emissions from a wetland soil exposed to elevated CO_2 . *Biogeochemistry* **37**, 77–88 (1997).
- Ziska, L. H. *et al.* Long-term growth at elevated carbon dioxide stimulates methane emission in tropical paddy rice. *Glob. Change Biol.* **4**, 657–665 (1998).
- Strom, L., Ekberg, A., Mastepanov, M. & Christensen, T. R. The effect of vascular plants on carbon turnover and methane emissions from a tundra wetland. *Glob. Change Biol.* **9**, 1185–1192 (2003).
- Waddington, J. M. & Roulet, N. T. Groundwater flow and dissolved carbon movement in a boreal peatland. *J. Hydrol.* **191**, 122–138 (1997).
- Kuzakov, Y. Review: Factors affecting rhizosphere priming effects. *J. Plant Nutr. Soil Sci.* **165**, 382–396 (2002).
- Wetzel, R. G. Gradient-dominated ecosystems—Sources and regulatory functions of dissolved organic matter in freshwater ecosystems. *Hydrobiologia* **229**, 181–198 (1992).

Acknowledgements We are grateful to the Royal Society, the Welsh Assembly Government and the Natural Environment Research Council, UK, for funding this research.

Competing interests statement The authors declare competing financial interests: details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to C.F. (c.freeman@bangor.ac.uk).

Unexpectedly recent dates for human remains from Vogelherd

Nicholas J. Conard¹, Pieter M. Grootes² & Fred H. Smith³

¹Abteilung für Ältere Urgeschichte und Quartärökologie Institut für Ur- und Frühgeschichte und Archäologie des Mittelalters Universität Tübingen, Schloss Hohentübingen, 72070 Tübingen, Germany

²Leibniz Labor für Altersbestimmung und Isotopenforschung Universität Kiel, Max-Eyth-Strasse 11–13 24118 Kiel, Germany

³Department of Anthropology, Loyola University Chicago, 6525 North Sheridan Road, Chicago, Illinois 60626, USA

The human skeletal remains from the Vogelherd cave in the Swabian Jura of southwestern Germany are at present seen as the best evidence that modern humans produced the artefacts of the early Aurignacian¹. Radiocarbon measurements from all the key fossils from Vogelherd show that these human remains actually date to the late Neolithic, between 3,900 and 5,000 radiocarbon years before present (BP). Although many questions remain unresolved, these results weaken the arguments for the Danube Corridor hypothesis²—that there was an early migration of modern humans into the Upper Danube drainage—and strengthen the view that Neanderthals may have contributed significantly to the development of Upper Palaeolithic cultural traits independent of the arrival of modern humans^{3,4}.

Since the discovery of anatomically modern skeletal remains with Aurignacian artefacts at the Cro-Magnon rock shelter in 1868, the Aurignacian, as this cultural group later became known, has been viewed as a product of modern humans⁵. The presence of Aurignacian artefacts has often been equated with modern humans even in settings lacking human skeletal material^{2,6,7}. Researchers have traditionally viewed the Cro-Magnon skeletal remains from southwestern France as the definitive association between modern humans and the early Aurignacian^{6,8}, but a recent AMS (accelerator mass spectrometer) date associated with these specimens places them within the early Gravettian at approximately 27,760 radiocarbon yr BP⁹. Scholars have suggested that other skeletal remains demonstrate an association between the early Aurignacian and modern humans, but virtually all of these specimens lack diagnostic morphology or are of dubious contextual association¹.

A possible exception is the site of Mladeč (Lautsch) in the Czech Republic, which was excavated in the late nineteenth and early twentieth centuries. Here modern human remains are purported to be associated with Aurignacian artefacts¹⁰. Recent dates of calcite deposits overlying the human bones yielded ages of 34,000–35,000 radiocarbon yr BP¹¹. However, owing to the complex depositional history of the site and problems associated with dating geological carbonate deposits, the age of the human skeletal remains will remain unknown until the human bones are dated directly.

Following the reassessment of the Cro-Magnon remains, the human specimens from the Vogelherd cave came to be viewed as the best evidence for an association between modern humans and early Aurignacian finds. Riek^{12,13} directed excavations at Vogelherd, near Stetten, in the Lone Valley of southwestern Germany in the summer and autumn of 1931 (Fig. 1). Over a period of three months the excavation team removed about 300 m³ of Pleistocene sediments from the cave and recovered artefacts from four Middle Palaeolithic and four Upper Palaeolithic layers. Riek published 12 stratigraphic profiles and detailed descriptions of the sediments from the cave. Aurignacian deposits from archaeological horizons (AH) V and IV are particularly noteworthy among the impressive archaeological deposits from the site¹². These deposits produced some of the richest Aurignacian assemblages in central Europe and included a dozen examples of early figurative art, scores of diverse

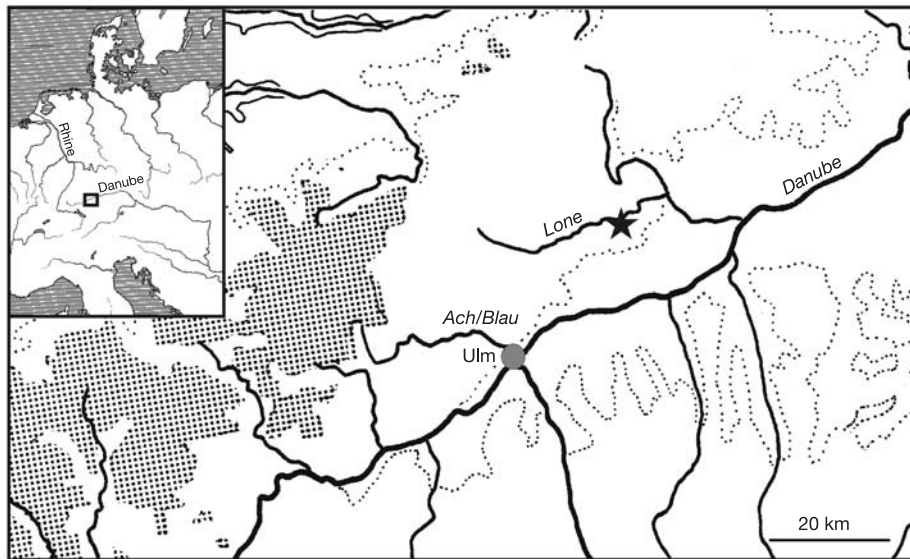


Figure 1 Map showing the location of Vogelherd cave. The star indicates the position of the site.

organic artefacts, and classically Aurignacian lithic artefacts dating to between 30,000 and 36,000 radiocarbon yr BP^{2,14}.

Riek and Gieseler reported well-preserved bones of modern humans, including a cranium and mandible (Stetten 1, see Fig. 2), a humerus (Stetten 3), two vertebrae (Stetten 4), a metacarpus (Stetten 5) from the base of the lower Aurignacian of AH V, and a cranium (Stetten 2) from the top of the upper Aurignacian of AH IV^{12,13,15}. With the exception of the humerus, which was recovered near the central part of the cave, and the Stetten 2 cranium, recovered near the southern entrance, the other finds were recovered near the southwest entrance. Riek was present on July 22 1931 when the Stetten 1 cranium was recovered, and described the provenance of this and the other fossils in great detail^{12,13}. His description of the stratigraphic position of the human remains is unambiguous and provides no indication that the finds might have been intrusive.

Thus human skeletal remains from Vogelherd constituted key evidence that modern humans produced these Aurignacian assemblages and by association the Aurignacian in general. Numerous morphological assessments of the site's cranial remains show them to be unquestionably modern^{1,15–17}. Whereas Gieseler¹⁵ suggested that the robust Vogelherd 3 humerus might represent a Neanderthal, systematic morphometric analysis has revealed the specimen's modern affinities¹⁸. A recent genetic study of the Vogelherd 3 humerus is consistent with this interpretation¹⁹. Direct radiocarbon dating was needed to confirm or refute the age attributed to the human skeletal remains on the basis of archaeological associations.

We selected five bones from the sites Stetten 1–4 for dating at the Leibniz Laboratory. After rigorous sample preparation, including soxhlet extraction with a suite of organic solvents to remove non-polar museum contaminants²⁰, and warm water extraction to remove bone glue, bone collagen was successfully isolated as filtered gelatin²¹ from all of the samples (Table 1). An organic residue on the silver filter provides a check on non-soluble contaminants. After combustion to CO₂ in quartz ampoules and reduction to graphite, the ¹⁴C concentrations of both fractions were measured by AMS²² and converted into radiocarbon ages following ref. 23, with a correction for isotope fractionation based on ¹³C/¹²C ratios measured simultaneously by AMS.

All collagen samples provided unexpectedly young, mid-Holocene ages. The reliability of these ages is supported by AMS results

on the non-soluble material, which produced ages within a few hundred years of the measurements on collagen. A yellow, resin-like material, which resisted our solvent extraction, was picked from one of the Stetten 4 vertebrae. This material has a ¹⁴C concentration of 135% of the modern standard, which corresponds to atmospheric ¹⁴C levels around AD 1977 and a brief period around 1962. Traces of this material in the non-soluble fraction may be responsible for three residue ages that are slightly younger than the collagen ages.

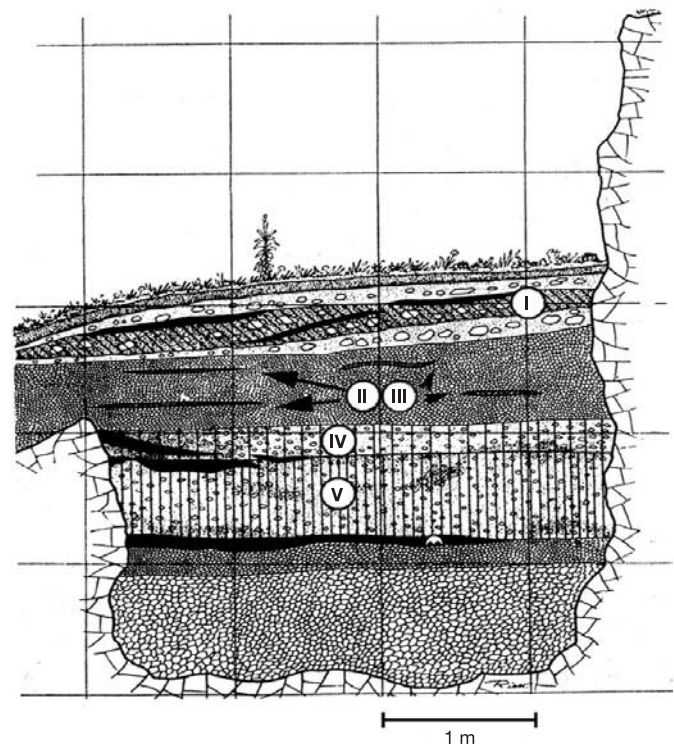


Figure 2 Stratigraphic location of the cranium and mandible (Stetten 1). I–V represent archaeological horizons I–V. As depicted, the Stetten 1 cranium underlies AH V. After Riek 1934.

Table 1 Radiocarbon ages of human skeletal remains from Vogelherd cave

Laboratory number	Specimen	Reported stratigraphic context	Collagen yield ¹ (mg)	Material	Date ³ (BP)
KIA 20967	Stetten 1 cranium	AH V base	2.9	Collagen	4,910 ± 25
			1.7	Insoluble residue	4,970 ± 35
KIA 20969	Stetten 1 mandible ²	AH V base	4.2	Collagen	4,985 ± 30
			0.7	Insoluble residue	5,070 ± 45
KIA 19538	Stetten 1 mandible	AH V base	3.8	Collagen	4,715 ± 35
			1.4	Insoluble residue	4,695 ± 35
KIA 19537	Stetten 2 cranium	AH IV top	3.5	Collagen	3,980 ± 35
			1.4	Insoluble residue	3,560 ± 30
KIA 19539	Stetten 4 vertebra	AH V base	1.3	Collagen	4,735 ± 30
			4.5	Insoluble residue	4,245 ± 25
			9.0	Organic preservative	135.1 ± 0.4 pMC ⁴ (AD 1962 or 1977)
KIA 19540	Stetten 3 humerus	AH V base	4.1	Collagen	4,995 ± 35
			2.2	Insoluble residue	5,175 ± 30

¹Collagen yields were all sufficient for reliable radiocarbon measurements. These ages should reflect the actual radiocarbon age of the human remains. The samples' weights for insoluble residues and organic preservatives are also listed in this column.

²The Stetten 1 mandible was dated twice with independent sample preparation.

³ ± 1 s.d.

⁴Per cent modern carbon (pMC).

The different age of the Stetten 2 cranium is not entirely surprising because it originates from a higher stratigraphic position ~30 m away from the concentration of human bones associated with the Stetten 1 cranium^{12,13}.

The six AMS radiocarbon dates on collagen from these skeletal remains, and their supporting dates, show that all of the Vogelherd human skeletal remains are from the Holocene and are irrelevant to discussions on the origin of modern Europeans (Table 1). We conclude that Riek's stratigraphic assessment of the fossils was incorrect. The human bones seem to originate from intrusive Neolithic burials near the southwestern and southern entrances to the cave. All 26 previously reported radiocarbon measurements on animal bones from Vogelherd have produced Pleistocene ages, with most dates within the expected range of the Aurignacian^{2,14}. Pleistocene fauna, including bones of reindeer, mammoth, woolly rhinoceros, horse and cave bear, comprise the vast majority of the assemblage, and there is no indication of a significant Holocene component at the site. This conclusion is entirely consistent with the presence of an extremely rich assemblage of Aurignacian artefacts¹².

These new dates raise many important questions about the critical period between ~40,000 and 30,000 yr BP, when both modern humans and Neanderthals occupied Europe^{1,2,24}. The Holocene dates for the Vogelherd human remains remove what has been the most convincing association of early Aurignacian assemblages with modern humans in Europe. Whereas there are other potential associations between modern humans and the Aurignacian, none of them provides particularly compelling cases, for varying reasons¹. Unfortunately, the important new skeletal material from Peștera cu Oase, Romania, recently dated at ~35,000 radiocarbon yr BP, lacks any archaeological association²⁵. Despite a lack of precise temporal resolution due to major fluctuations in levels of atmospheric radiocarbon^{2,26–29}, the new Romanian specimen strongly suggests that modern humans were present in Europe during the early Aurignacian, but this is also true for late Neanderthals^{24,30}.

The Holocene age of the human skeletal remains from Vogelherd places the question of who made the earliest Aurignacian in Europe in doubt. At present the hypothesis that the Neanderthals gave rise to the early Aurignacian, as has been argued by some colleagues including Richter³, cannot be refuted. Additionally, the Danube Corridor model for the early colonization of central Europe by modern humans², although still plausible, can no longer be demonstrated on the basis of associations between modern humans and the early Aurignacian at Vogelherd. With the new dates from Vogelherd one of the most widely held assumptions of paleo-anthropology—that the Aurignacian is uniquely associated with modern humans—seems more uncertain than ever. These results also create the possibility that the figurative art found at Vogelherd

was produced by Neanderthals. New excavations providing unequivocal associations between human skeletal remains and the early Aurignacian will be necessary to address these issues. □

Received 28 February; accepted 19 May 2004; doi:10.1038/nature02690.

- Churchill, S. E. & Smith, F. H. Makers of the Early Aurignacian of Europe. *Yb. Phys. Anthropol.* **43**, 61–115 (2001).
- Conard, N. J. & Bolus, M. Radiocarbon dating the appearance of modern humans and the timing of cultural innovations in Europe: new results and new challenges. *J. Hum. Evol.* **44**, 331–371 (2003).
- Richter, J. 'Out of Africa II' Die Theorie über die Einwanderung des modernen Menschen nach Europa auf dem archäologischen Prüfstand. *Archäol. Inform.* **19**, 67–73 (1996).
- D'Errico, F. The invisible frontier. A multiple species model for the origin of behavioral modernity. *Evol. Anthropol.* **12**, 188–202 (2003).
- Boule, M. *Les Hommes Fossiles* (Mason et Cie, Paris, 1921).
- Stringer, C. & Gamble, C. *Search of the Neanderthals* (Thames & Hudson, London, 1993).
- Klein, R. *The Human Career*, 2nd edn (Univ. Chicago Press, Chicago, 1999).
- Stringer, C., Hublin, J. J. & Vandermeersch, B. in *The Origins of Modern Humans* (eds Smith, F. & Spencer, F.) 51–135 (Alan Liss, New York, 1984).
- Henry-Gambier, D. Les fossiles de Cro-Magnon (Les Eyzies-de-Tayac, Dordogne): nouvelles données sur leur position chronologique et leur attribution culturelle. *Bull. Mém. Soc. Anthropol. Paris* **14**, 89–112 (2002).
- Svoboda, J. The depositional context of the Early Upper Paleolithic human fossils from the Koněprusy (Zlatý kůň) and Mladeč caves. Czech Republic. *J. Hum. Evol.* **38**, 523–536 (2000).
- Svoboda, J., van der Plicht, J. & Kuželka, V. Upper Paleolithic and Mesolithic human fossils from Moravia and Bohemia (Czech Republic): some new ¹⁴C dates. *Antiquity* **76**, 957–962 (2002).
- Riek, G. *Die Eiszeitjägerstation am Vogelherd im Lonetal I* (Heine, Tübingen, 1934).
- Riek, G. Paläolithische Station mit Tierplastiken und menschlichen Skelettresten bei Stetten ob Lontal. *Germania* **16**, 1–8 (1932).
- Conard, N. J., Niven, L., Mueller, K. & Stuart, A. The chronostratigraphy of the Upper Paleolithic of Vogelherd. *Mitt. Ges. Urgesch.* **12**, 73–86 (2003).
- Gieseler, W. Bericht über die jungpaläolithischen Skelettreste von Stetten ob Lontal bei Ulm. *Verh. Ges. Phys. Anthropol.* **8**, 41–48 (1937).
- Gieseler, W. Die urgeschichtlichen Menschen-Funde aus dem Lonetal und ihre Bedeutung für die deutsche Urgeschichte. *Jahresbande Wiss. Akad. Tübingen NCD Dozentbundes* **1**, 102–127 (1940).
- Czarnetzki, A. in *Urgeschichte in Baden-Württemberg* (ed. Müller-Beck, H.) 217–240 (Konrad Theiss, Stuttgart, 1983).
- Churchill, S. E. & Smith, F. H. A modern human humerus from the early Aurignacian of Vogelherdhöhle (Stetten, Germany). *Am. J. Phys. Anthropol.* **112**, 251–273 (2000).
- Scholz, M. et al. Genomic differentiation of Neanderthals and anatomically modern man allows a fossil-DNA-based classification of morphologically indistinguishable hominid bones. *Am. J. Hum. Genet.* **66**, 1927–1932 (2000).
- Bruh, F., Dühr, A., Grootes, P. M., Mintrop, A. & Nadeau, M.-J. Chemical removal of conservation substances by 'Soxhlet'-type extraction. *Radiocarbon* **43**, 229–237 (2001).
- Grootes, P. M., Nadeau, M. J. & Rieck, A. ¹⁴C AMS at the Leibniz-Labor: radiocarbon dating and isotope research. *Nuclear Instruments and Methods* (in the press).
- Nadeau, M. J. et al. The Leibniz-Labor AMS facility at the Christian-Albrechts-University, Kiel, Germany. *Nucl. Instrum. Methods B* **123**, 22–30 (1997).
- Stuiver, M. & Polach, H. Discussion: reporting of ¹⁴C data. *Radiocarbon* **19**, 355–363 (1977).
- Smith, F. H., Trinkaus, E., Pettitt, P. B., Karavanić, I. & Paunović, M. Direct radiocarbon dates for Vindija G₁ and Velika Pečina Late Pleistocene hominid remains. *Proc. Natl Acad. Sci. USA* **96**, 12281–12286 (1999).
- Trinkaus, E. et al. An early modern human from the Peștera cu Oase, Romania. *Proc. Natl Acad. Sci. USA* **100**, 11231–11236 (2003).
- Beck, J. W. et al. Extremely large variations of atmospheric ¹⁴C concentration during the last glacial period. *Science* **292**, 2453–2458 (2001).
- Hughen, K. A. et al. ¹⁴C activity and global carbon cycle changes over the past 50,000 years. *Science* **303**, 202–207 (2004).
- Voelker, A. H. L., Grootes, P. M., Nadeau, M. J. & Sarnthein, M. Radiocarbon levels in the Iceland Sea from 25–53 kyr and their link to the earth's magnetic field intensity. *Radiocarbon* **42**, 437–452 (2000).

29. Van Kreveld, S. *et al.* Potential links between surging ice sheets, circulation changes, and the Dansgaard-Oeschger cycles in the Irminger Sea, 60–18 kyr. *Paleoceanography* **15**, 425–442 (2000).
30. Hublin, J. J., Barroso Ruiz, C., Medina Lara, P., Fontugne, M. & Reyss, J.-L. The Mousterian site of Zafarraya (Andalucía, Spain): dating and implications on the palaeolithic peopling processes of Western Europe. *C.R. Acad. Sci.* **321**, 931–937 (1995).

Acknowledgements We are grateful to M. Bolus, C. Pusch, H. Floss, M. Haidle, M. Malina, L. Niven and E. Trinkaus for their assistance and discussions, and we thank the Leibniz team for cleaning and dating the bones. This work was funded by the Landesdenkmalamt Baden-Württemberg, the Alb-Donau-Kreis, the Deutsche Forschungsgemeinschaft, the Alexander von Humboldt Stiftung, and the Eberhard-Karls-Universität Tübingen.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.J.C. (nicholas.conard@uni-tuebingen.de).

Origin of extant domesticated sunflowers in eastern North America

Abigail V. Harter¹, Keith A. Gardner¹, Daniel Falush², David L. Lentz³, Robert A. Bye⁴ & Loren H. Rieseberg¹

¹Department of Biology, Indiana University, Bloomington, Indiana 47405, USA

²Peter Medawar Building for Pathogen Research, University of Oxford, South Parks Road, Oxford OX1 3SY, UK

³Chicago Botanic Garden, 1000 Lake Cook Road, Glencoe, Illinois 60022, USA

⁴Jardín Botánico Exterior, Instituto de Biología, UNAM, México DF 04510, México

Eastern North America is one of at least six regions of the world where agriculture is thought to have arisen wholly independently^{1–5}. The primary evidence for this hypothesis derives from morphological changes in the archaeobotanical record of three important crops—squash, goosefoot and sunflower—as well as an extinct minor cultigen, sumpweed^{1,3}. However, the geographical origins of two of the three primary domesticates—squash and goosefoot—are now debated^{6,7}, and until recently sunflower (*Helianthus annuus* L.) has been considered the only undisputed eastern North American domesticate. The discovery of 4,000-year-old domesticated sunflower remains from San Andrés, Tabasco^{8,9}, implies an earlier and possibly independent origin of domestication in Mexico and has stimulated a re-examination of the geographical origin of domesticated sunflower. Here we describe the genetic relationships and pattern of genetic drift between extant domesticated strains and wild populations collected from throughout the USA and Mexico. We show that extant domesticates arose in eastern North America, with a substantial genetic bottleneck¹⁰ occurring during domestication.

There are two hypotheses regarding the origin of agriculture in eastern North America. One hypothesis holds that agriculture arose independently in this region with the domestication of four to seven indigenous species^{1–3,8,11}. The alternative states that most major cultigens originated in Mesoamerica and were dispersed northwards to the eastern woodlands of North America, triggering the domestication of minor indigenous crops^{1–3,8,11}.

Before the discovery of domesticated sunflower remains in Mexico, the sunflower provided the most convincing evidence for an independent origin of agriculture in eastern North America. Although the progenitor of domesticated sunflower, *H. annuus*, is now distributed across North America from southern Canada to northern Mexico¹², wild populations from the east-central USA are most similar morphologically to the domesticates, and domesticated sunflower remains are found at several archaeological sites in

this region. Thus, early authors placed the origin of domestication in eastern North America^{1,3,11–16} and identified the east-central wild form as a probable progenitor^{13–16}. Molecular genetic studies completed before the Tabasco, Mexico, discovery were inconclusive regarding both the number of origins and geographical source of the domesticated sunflower, but in these studies wild populations were insufficiently sampled^{10,17–19} and the molecular markers employed were insufficiently variable to resolve genetic relationships^{17–19}.

To determine the geographical origin(s) of sunflower domestication and to account for the genetic composition of extant domesticates, we have used model-based methods to evaluate genetic relationships and reconstruct the pattern of genetic drift among 21 populations of wild *H. annuus* and eight Native American landraces from the USA and Mexico, as well as two modern cultivars (USDA and Mammoth) (Fig. 1, Supplementary Table 1). The results described are based upon data from 18 microsatellite loci distributed across the sunflower genome (Supplementary Table 2).

To identify ancestral source populations for extant domesticates we used the ‘admixture model’ included in the software program STRUCTURE^{20,21} to infer population structure in wild *H. annuus* and assign domesticates to inferred populations. In this bayesian approach, multilocus genotypic data are used to define a set of populations with distinct allele frequencies, hereafter referred to as clusters, and assign individuals probabilistically to these defined clusters with or without prior knowledge of sampling location. Also, the admixture model assumes that loci are unlinked and can freely recombine.

Without specifying prior information concerning sampling location, and allowing for admixed individuals, we estimated the number of genetic clusters of wild *H. annuus* to which we would assign the domesticates (Methods, Supplementary Methods). Combining the results from these tests with geography, we modelled the assignment on two scales: regional and local. Using our estimate of population structure on a regional scale, we defined all Mexican populations plus Arizona as one potential source cluster and central US populations as a second potential source cluster (Supplementary Fig. 1). Although alleles are widespread across both regions and there are no significant differences in heterozygosity ($P > 0.42$, two-sided) or allelic richness ($P > 0.18$, two-sided) between the two regions and each domesticated individual was allowed to have originated from more than one source, all domesticates were assigned to the US cluster in all ten runs of the algorithm. The proportion of each domesticated individual’s genome having ancestry in the USA was ≥ 0.985 for all individuals, and for each domesticated strain, the average estimated ancestry in the US cluster was ≥ 0.997 (Fig. 2).

Again using the results of our analysis of population structure in wild *H. annuus*, we further modelled the assignment of the domesticates by subdividing the regional groups into four local area source clusters corresponding to west Mexico, east-central Mexico, the US Great Plains and the east-central USA (Supplementary Fig. 2). In all ten runs of the algorithm, again allowing for admixed origins, all domesticates were assigned to the east-central USA. The proportion of each individual’s genome having ancestry from this area was ≥ 0.896 , and for each domesticated strain, average estimated ancestry in the east-central USA was ≥ 0.994 (Fig. 2). Thus, the results of both our ancestry analyses indicate that the ten strains of domesticated *H. annuus* are genetically most similar to wild populations from the central USA, particularly the easternmost populations in our sample.

When considered as a group, the genetic diversity in domesticated sunflower is significantly less than the genetic diversity in wild *H. annuus* (Supplementary Table 3). We hypothesized that the domesticates’ low genetic diversity is a consequence of strong genetic drift from central US populations, a scenario compatible with bottlenecks that would have occurred owing to strong selection during domestication¹⁰. To investigate the historical processes

underlying this pattern of reduced diversity, we compared the rate of drift away from a common ancestor for each wild population and domesticated strain, using the 'F model' as implemented with the program STRUCTURE²¹. The model assumes that each population in the comparison has undergone independent drift away from the allele frequencies found in their common ancestor. A bayesian approach is used to make inferences about ancestral allele frequencies as well as the rates of drift away from the ancestral allelic state in each population, hereafter referred to as *F* values. In wild-domestic comparisons, wild populations whose allele frequencies most closely resemble those found in the common ancestor of wild and domesticated *H. annuus* should exhibit little drift away from the ancestral state, that is, they should have low *F* values. Furthermore, if domestication was associated with a genetic bottleneck, drift in the domesticated strains should be significantly higher than drift in the wild populations.

For each pairwise comparison of wild population and domesticated strain, the estimated *F* values were consistent across three runs of the model and a clear geographically correlated pattern of drift was detected in the wild populations (Supplementary Table 4). The 90% credibility regions around the estimated *F* values for four wild populations—Kansas, South Dakota, Oklahoma2 and Iowa—

include zero drift away from the ancestral allele frequencies across all comparisons. These results are consistent with the genetic composition of these modern populations being nearly identical to that of the wild progenitors of domesticated sunflowers.

The drift values estimated for the domesticates are much higher than those found in the wild populations and consistent with strong genetic drift during domestication (Supplementary Table 4). The lowest mean value for any domesticate (Havasupai) is 0.2864, which is 220 times the lowest mean drift value for any wild population (0.0014) (Fig. 3). The confidence limits for these values are relatively narrow, implying that the genetic composition of the domesticates has changed at least 50-fold faster than the wild populations since they diverged. Furthermore, the variation in extant genetic diversity among domesticated lineages suggests that multiple founder events and bottlenecks¹⁰ subsequent to the initial domestication event are responsible (Supplementary Table 3). Again, this scenario is borne out in our genetic drift analysis: although drift in all domesticates is very high, it ranges from around 0.3 for Arikara, Havasupai, Maíz de Tejas, Mammoth and Seneca to around 0.5 for Hopi, Mandan and USDA.

The *F* model results are consistent with the ancestry analysis presented above, supporting a central US origin for domesticated

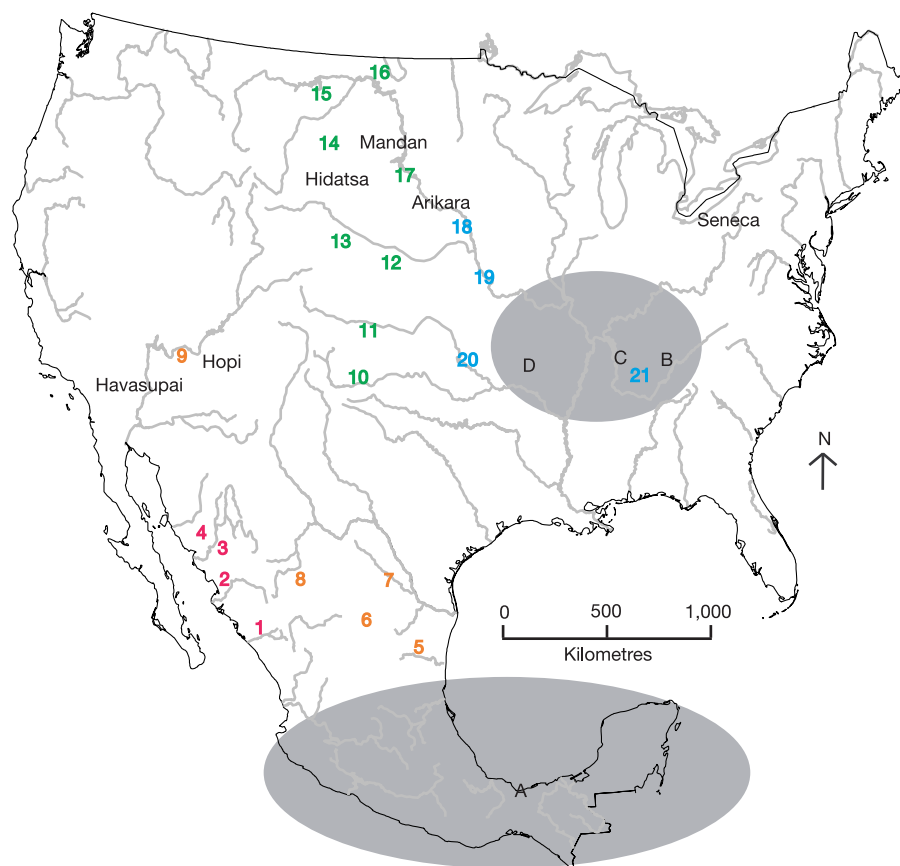


Figure 1 Map of sampling locations, archaeological sites and Native American groups. Shaded areas indicate centres of domestication with eastern North America¹ to the north and Mesoamerica (as defined in ref. 26) to the south. Numbers indicate sampling locations of wild populations where 1 is Sinaloa, 2 is Sonora5, 3 is Sonora4, 4 is Sonora6, 5 is Tamaulipas, 6 is Zacatecas, 7 is Nuevo León, 8 is Chihuahua, 9 is Arizona, 10 is Texas, 11 is Oklahoma2, 12 is Kansas, 13 is Colorado, 14 is Montana1, 15 is Montana2, 16 is North Dakota, 17 is South Dakota, 18 is Iowa, 19 is Missouri, 20 is Oklahoma1 and 21 is Tennessee (colours correspond to local area model ancestry analysis). Letters indicate archaeological sites with the oldest remains of domesticated sunflower, where A

is San Andrés, Tabasco, Mexico ($4,130 \pm 40$ radiocarbon years before present (BP)), B is Higgs, Tennessee, USA ($2,850 \pm 85$ radiocarbon years BP), C is Hayes, Tennessee, USA ($4,265 \pm 60$ radiocarbon years BP) and D is Marble Bluff, Arkansas, USA ($2,843 \pm 44$ radiocarbon years BP); and names indicate the historical locations of Native American groups. We note that although they were collected in Mexico, the identities of indigenous groups associated with Maíz de Tejas and Maíz negro are unknown. USDA and Mammoth are modern cultivars derived from Russian stock. Therefore, these strains do not appear on the map.

sunflowers (Fig. 3). The nine lowest mean F values are all for central US populations. Seven out of the eight Mexican populations including Arizona have mean F values many times higher than those found in this group of nine central US populations, and their 90% credibility regions do not overlap with this group in any of the domesticated–wild comparisons (Supplementary Table 4). Only the northeasternmost Mexican population (Nuevo León) has an intermediate mean F value; an analysis of ancestry using the admixture model estimated 70% Mexican and 30% US ancestry for this population, indicating that its intermediate mean F value is caused by genetic exchange with wild central US populations. Three of the central US populations—North Dakota, Oklahoma1 and Tennessee—have diverged from the population ancestral to domesticated sunflowers, as indicated by intermediate or high mean F values relative to other wild populations in the central USA. These populations nevertheless cluster with other US populations in our analysis of population structure in wild *H. annuus* and on the neighbour-joining tree, which is consistent with a localized drift event in these locations.

Given the substantial genetic bottleneck common to all domesticated sunflowers and the evidence for genetic fluidity among individual wild populations, it may be that genetic evidence alone

will be of limited value in further narrowing down the site or region of origin of modern domesticated sunflowers, so that a multi-disciplinary approach may be required. Archaeological evidence of sunflower domestication has been reported at multiple sites in the region extending from the Appalachian west wall to the prairie margin in the modern-day states of Tennessee, Kentucky, Iowa, Missouri, Arkansas and Ohio^{11,15,22,23}. Historical records suggest that wild *H. annuus* was introduced by humans to this area from the adjacent central plains and subsequently brought under cultivation¹⁴. Geographically distant indigenous groups have developed extant Native American landraces for the seed as a source of food and a source of oil or for the achene coat as a source of dye¹³.

We have shown that genetic relationships and historical drift between wild populations and extant domesticates support the hypothesis that extant domesticated sunflowers arose from wild populations in the central USA via strong genetic drift as would be expected from a selective bottleneck during domestication in eastern North America. Thus, our findings are consistent with earlier archaeological and morphological conclusions and agree with current research from other continents indicating that the transition to agriculture involves complex regional models rather than simple diffusion^{2–4}. The possibility of an earlier and independent domes-

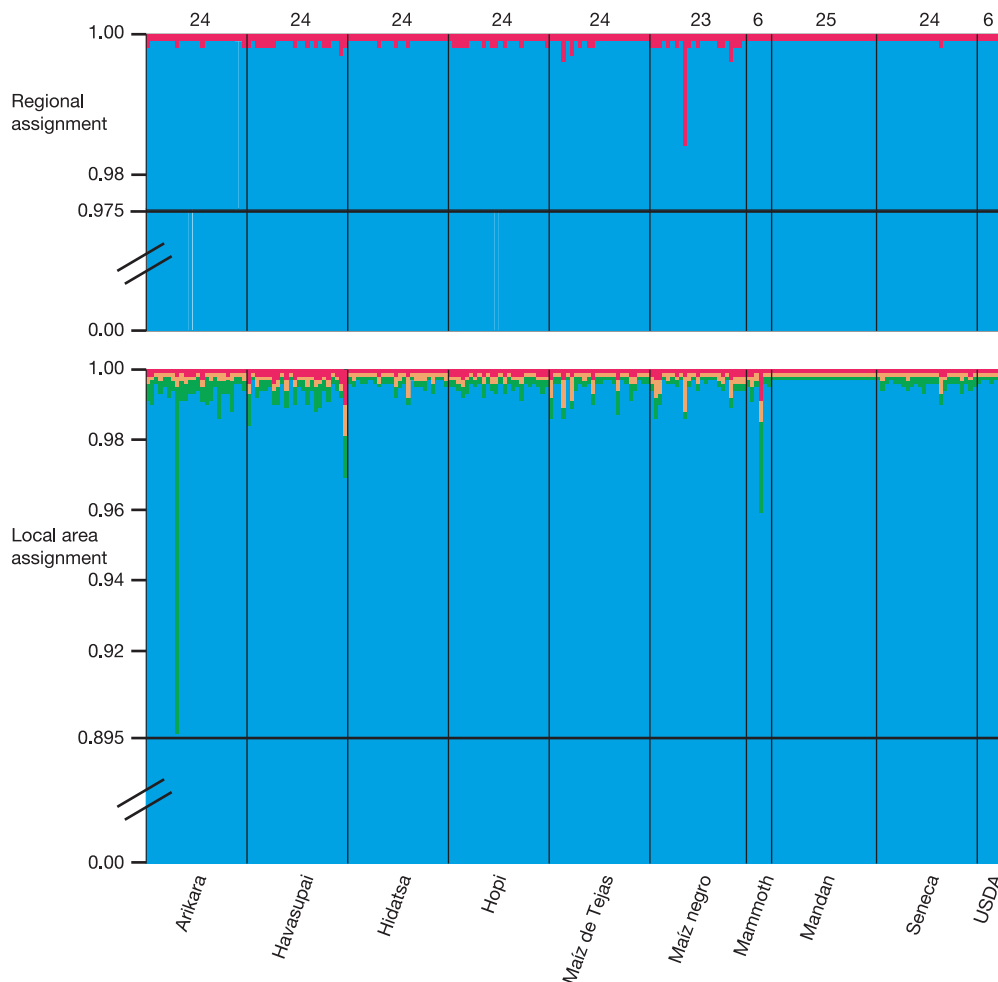


Figure 2 Results of domesticated *H. annuus* ancestry analysis. Each domesticated individual's genome is represented by a thin vertical line partitioned into coloured segments in proportion to the estimated ancestry in each source cluster. Strains are separated by black lines, with names below and sample sizes above. Results are shown

for one run each of: regional model (upper frame) with Mexico plus Arizona (red) and central USA (blue), local area model (lower frame) with west Mexico (red), east-central Mexico (orange), US Great Plains (green), east-central USA (blue). Each run of the algorithm for each model yielded nearly identical results.

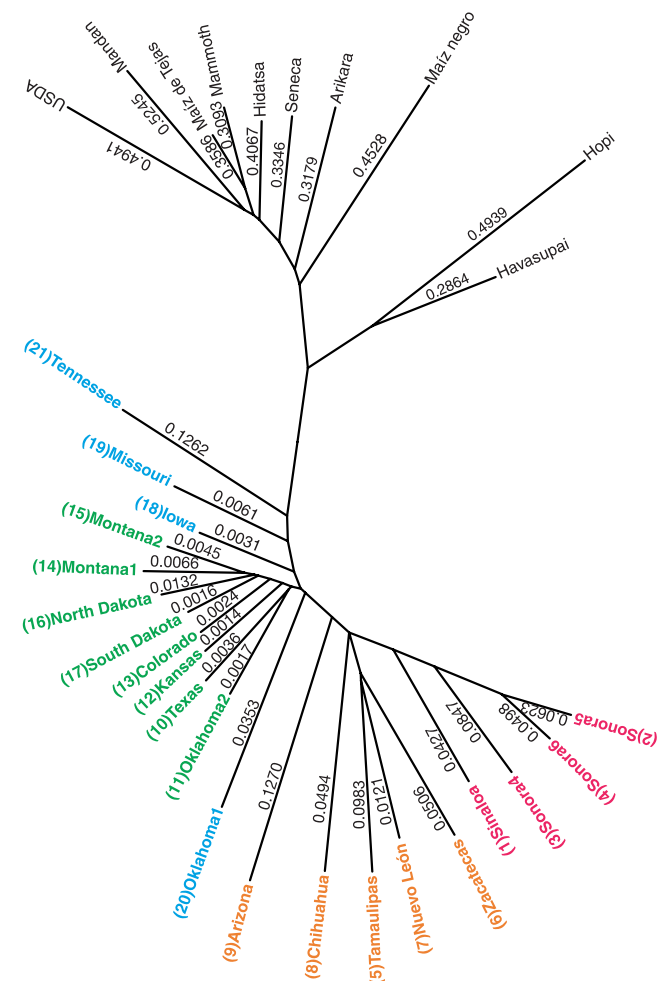


Figure 3 Comparisons of genetic drift in wild populations (colours correspond to local area model source clusters) and domesticated strains (in black). The lines represent a neighbour-joining tree summarizing the genetic distances, D_A , between groups. Mean F values for each domesticated strain averaged across all wild comparisons and mean F values for each wild population averaged across all domesticate comparisons appear along lines. Numbers in brackets correspond to those on Fig. 1 for sampling locations.

tication event in Mexico, as suggested by the discovery in Tabasco, is not ruled out. However, our results indicate that a Mexican domesticate did not contribute to extant domesticated germplasm. To determine the significance of the Mexican palaeoethnobotanical sunflower findings, further archaeological work elucidating the origin and fate of the domesticated sunflower in Mexico is strongly recommended. □

Methods

Sampling and data collection

For DNA extraction, leaf tissue was sampled in the field from twenty-one wild populations of *H. annuus* and from greenhouse plants grown from seed, representing eight Native American landraces (Arikara, Havasupai, Hidatsa, Hopi, Maiz de Tejas, Maiz negro, Mandan and Seneca) and two modern cultivars (USDA, Mammoth) (Fig. 1 and Supplementary Table 1). Microsatellite loci were amplified by polymerase chain reaction (PCR) with fluorescence-labelled primers and surveyed in all sampled wild populations and domesticates (Supplementary Table 2). Of these 21 loci, data for three were excluded from further analysis owing to an apparent excess of homozygosity, presumably due to null alleles. PCR amplifications were carried out according to standard protocol and electrophoresis was carried out on an ABI Prism 3700 DNA analyser (PE Biosystems). Peak data were analysed using Genescan 3.5 and scoring of markers was performed with Genotyper 3.6NT (PE Biosystems programs).

Genetic diversity estimation

Gene diversity (average heterozygosity) in each wild population and each domesticated strain was estimated using DISPAN²⁴. Diversity measures were calculated and comparison tests were performed using FSTAT version 2.9.3 (ref. 25; Supplementary Table 3). P values were estimated using a permutation method.

Inferring population structure in wild populations of *H. annuus*

Using the admixture model^{20,21} (STRUCTURE version 2) we estimated the number of genetic clusters, K , to which we would assign the domesticated individuals without specifying prior information concerning sampling location. Using the full data set, three to ten parallel Markov chains were run for all models of K with a burn-in of 50,000 iterations and a run length of 10^6 iterations following the burn-in. For each run, the ln likelihood of each model was calculated. The full data set was analysed for all models from $K = 1$ through to 12 and also for $K = 21$ clusters (21 is the total number of sampled populations). For all $K = 21$ runs, individuals were assigned to no more than 12 clusters. Therefore, in additional analyses we tested only models of $K = 1$ through to 12. For the full data set, a single splitting solution was found for $K = 2$ in which the sampled populations were clustered into two geographical regions: all Mexican populations plus Arizona versus central US populations (Supplementary Fig. 1). For further analysis the data set was subdivided into the Mexico plus Arizona subsample and the central US subsample (Supplementary Methods). The former was tested for $K = 1$ through to 9 and the latter was tested for $K = 1$ through to 12 running the algorithm three to five times for each K . At $K \geq 2$, the central USA was split into two clusters corresponding to the upper Great Plains and the east-central USA (Supplementary Fig. 2). For the Mexico plus Arizona region, $K = 2$ yielded two solutions that divided the sampled populations into a west coastal area and an east-central area (Supplementary Fig. 2).

Ancestry analysis of the domesticated *H. annuus* individuals

Using STRUCTURE version 2, ten parallel Markov chains were run for each model with a burn-in of 30,000 iterations and a run-length of 10^5 iterations following the burn-in. All models allowed for admixed individuals (Supplementary Methods). For each run, wild individuals were specified as belonging to pre-determined source clusters, but no prior information was specified as to the origin of domesticated individuals. We estimated the ancestry in each source cluster averaged across all individuals per domesticated strain and the proportion of each domesticated individual's genome having ancestry in each cluster.

Genetic drift analysis

For all pairs of wild populations and domesticated strains, we used the F model²¹ (STRUCTURE version 2) to estimate the amount of drift that has occurred in each lineage since their divergence. We specified in advance the origin of each individual as either a domesticate or wild with a prior probability of misassignment to these categories of 0.001. We set the mean and variance of the prior on F to 0.1, which puts approximately equal weight on low and high values. Three parallel Markov chains were run for each comparison with a burn-in of 50,000 iterations and a run length of 10^6 iterations following the burn-in. Regions of 90% credibility were calculated from the distribution of F values estimated during the final run (Supplementary Table 4). Mean drift values for each wild population were calculated across all runs and all domesticates. Mean drift values for each domesticated strain were calculated across all runs and all wild populations.

Received 11 November 2003; accepted 1 June 2004; doi:10.1038/nature02710.

- Smith, B. D. Origins of agriculture in eastern North America. *Science* **246**, 1566–1571 (1989).
- Cowan, C. W. & Watson, P. J. in *The Origins of Agriculture: An International Perspective* (eds Cowan, C. W. & Watson, P. J.) 207–212 (Smithsonian Institution Press, Washington DC/London, 1992).
- Smith, B. D. *Rivers of Change: Essays on Early Agriculture in Eastern North America* (Smithsonian Institution Press, Washington DC/London, 2002).
- Diamond, J. Evolution, consequences and future of plant and animal domestication. *Nature* **418**, 700–707 (2002).
- Neumann, K. New Guinea: A cradle of agriculture. *Science* **301**, 180–181 (2003).
- Wilson, H. D. *Quinnia* and relatives (*Chenopodium* sect. *Chenopodium* subsect. *Cellulata*). *Econ. Bot.* **44S**, 92–110 (1990).
- Sanjurjo, O. I., Piperno, D. R., Andres, T. C. & Wessel-Beaver, L. Phylogenetic relationships among domesticated and wild species of *Cucurbita* (Cucurbitaceae) inferred from a mitochondrial gene: Implications for crop plant evolution and areas of origin. *Proc. Natl Acad. Sci. USA* **99**, 535–540 (2002).
- Lentz, D. L., Pohl, M. E. D., Pope, K. O. & Wyatt, A. R. Prehistoric sunflower (*Helianthus annuus* L.) domestication in Mexico. *Econ. Bot.* **55**, 370–376 (2001).
- Pope, K. O. *et al.* Origin and environmental setting of ancient agriculture in the lowlands of Mesoamerica. *Science* **292**, 1370–1373 (2001).
- Tang, S. X. & Knapp, S. J. Microsatellites uncover extraordinary diversity in native American land races and wild populations of cultivated sunflower. *Theor. Appl. Genet.* **106**, 990–1003 (2003).
- Fritz, G. J. in *People, Plants and Landscapes* (ed. Gremillion, K. J.) 42–62 (Univ. Alabama Press, Tuscaloosa, Alabama, 1997).
- Heiser, C. B. Jr, Smith, D. M., Clevenger, S. B. & Martin, W. C. Jr The North American sunflowers (*Helianthus*). *Mem. Torr. Bot. Club* **22**, 1–218 (1969).
- Heiser, C. B. Jr The sunflower among the North American Indians. *Proc. Am. Phil. Soc.* **95**, 432–448 (1951).
- Asch, D. L. Common sunflower (*Helianthus annuus* L.): the pathway toward its domestication. *Proc. 58th Annu. Meeting 17 May 1993* (Society of American Archaeology, St Louis, Missouri).
- Yarnell, R. A. in *The Nature and Status of Ethnobotany in Eastern North America* (ed. Ford, R. I.) 289–300 (Anthropological Papers No. 67, Museum of Anthropology, Univ. Michigan, Ann Arbor, 1993).

16. Heiser, C. B. Jr Variation and subspeciation in the common sunflower, *Helianthus annuus*. *Am. Midl. Nat.* **51**, 287–305 (1954).
17. Rieseberg, L. H. & Seiler, G. Molecular evidence and the origin and development of the domesticated sunflower (*Helianthus annuus*). *Econ. Bot.* **44S**, 79–91 (1990).
18. Arias, D. M. & Rieseberg, L. H. Genetic relationships among domesticated and wild sunflowers (*Helianthus annuus*, Asteraceae). *Econ. Bot.* **49**, 239–248 (1995).
19. Cronn, R., Brothers, M., Klier, K., Bretting, P. K. & Wendel, J. F. Allozyme variation in domesticated annual sunflower and its wild relatives. *Theor. Appl. Genet.* **95**, 532–545 (1997).
20. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
21. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
22. Crites, G. D. Domesticated sunflower in fifth millennium B.P. temporal context: new evidence from middle Tennessee. *Tenn. Am. Antiquity* **58**, 146–148 (1993).
23. Brewer, A. J. in *Excavation of the Higgs and Doughty Sites I-25 Salvage Archaeology* (eds McCollough, M. C. R. & Faulkner, C. H.) 141–144 (Miscellaneous Papers No. 12, Tennessee Archaeological Society, Knoxville, Tennessee, 1973).
24. Ota, T. *DISPAN: Genetic Distance and Phylogenetic Analysis* (<http://www.bio.psu.edu/People/Faculty/Nei/Lab/Programs.html>) (Institute of Molecular Evolutionary Genetics, Pennsylvania State Univ., 1993).
25. Goudet, J. FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3), 2001. Updated from Goudet, J. FSTAT (version 1.2): a computer program to calculate F-statistics. *J. Hered.* **86**, 485–486 (1995).
26. McClung De Tapia, E. in *The Origins of Agriculture: An International Perspective* (eds Cowan, C. W. & Watson, P. J.) 144–145 (Smithsonian Institution Press, Washington DC/London, 1992).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank C. B. Heiser, Jr for advice throughout the course of this study and for reading an earlier version of the manuscript. We thank J. Pritchard for advice concerning implementation of the admixture model and A. Hwang and J. Murphy for assistance with the genotyping of the domesticates and Mexican populations, respectively. The analyses were made possible via access to the UNIX computing cluster provided by the University Information Technology Services department, Indiana University. This work was supported by grants from the NSF, the USDA and the National Geographic Society.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.V.H. (avharter@bio.indiana.edu).

Evolutionary change from induced to constitutive expression of an indirect plant resistance

Martin Heil¹, Sabine Greiner^{1,2}, Harald Meimberg³, Ralf Krüger¹, Jean-Louis Noyer¹, Günther Heubl³, K. Eduard Linzenmair² & Wilhelm Boland¹

¹Department of Bioorganic Chemistry, Max-Planck-Institute for Chemical Ecology, Hans-Knöll-Straße 8, Beutenberg Campus, D-07745 Jena, Germany

²Lehrstuhl Zoologie III, Biozentrum, Am Hubland, D-94074 Würzburg, Germany

³Department of Biology I, Section: Biodiversity Research, Systematic Botany, LMU Munich, Menzinger Straße 67, D-80638 München, Germany

⁴CIRAD, UMR 1096/PIA, TA 40/03, Avenue Agropolis, F-34398 Montpellier Cedex 5, France

Induced plant resistance traits are expressed in response to attack and occur throughout the plant kingdom^{1,2}. Despite their general occurrence, the evolution of such resistances has rarely been investigated³. Here we report that extrafloral nectar, a usually inducible trait, is constitutively secreted by Central American *Acacia* species that are obligately inhabited by ants. Extrafloral nectar is secreted as an indirect resistance⁴, attracting ants that defend plants against herbivores⁵. Leaf damage induces extrafloral nectar secretion in several plant species^{6–8}; among these are various *Acacia* species and other Fabaceae investigated here. In contrast, *Acacia* species obligately inhabited by symbiotic ants⁹

nourish these ants by secreting extrafloral nectar constitutively at high rates that are not affected by leaf damage. The phylogeny of the genus *Acacia* and closely related genera indicate that the inducibility of extrafloral nectar is the plesiomorphic or ‘original’ state, whereas the constitutive extrafloral nectar flow is derived within *Acacia*. A constitutive resistance trait has evolved from an inducible one, obviously in response to particular functional demands.

Induced resistances to herbivores have been described for more than 100 plant species² and can greatly benefit plants^{7,10–13}. They are generally regulated by the octadecanoid pathway, in which the plant hormone jasmonic acid forms a central signal^{2,14,15}. This pathway is taxonomically widely distributed and thus has to be regarded as evolutionarily highly conserved. Therefore, the question arises as to whether the expression regime of induced resistance traits can be evolutionarily adapted to particular functional demands.

We used extrafloral nectar (EFN) secretion by Central American *Acacia* (Fabaceae subfamily Mimosoideae) species to investigate whether a resistance trait can be differently expressed in closely related species. EFN is secreted by all species through glands on the leaf stalks, but it serves two different functions. The obligate myrmecophytes among these species permanently house ant colonies in their hollow thorns⁹. These symbiotic ants defend their host against herbivores and competing vegetation. They are nourished by plant-derived cellular protein-rich food bodies¹⁶ and by EFN, and both the ants and the plants seem to be highly adapted to this mutualism^{9,17}. Other, non-myrmecophytic species secrete EFN that is consumed by non-specialized ants from the vicinity. Agrawal & Rutter¹⁸ stated that ants obligately inhabiting myrmecophytes, in general recruit actively to parts of their host that currently require defence. In contrast, attraction of unspecialized ants to non-myrmecophytes was predicted to be controlled by the plants, for example, by short-term increases in the provisioning of food rewards. EFN secreted by *Acacia* functions as a regular food source for specialized mutualists of myrmecophytes and as a ‘bait’ attracting ants facultatively to non-myrmecophytes, thus allowing a test of these predictions.

We studied five myrmecophytic and four non-myrmecophytic *Acacia* species (all of the subgenus *Acacia*) as well as three species of related genera of the Mimosoideae. Study sites were located at the Pacific coast and in the Isthmus of Tehuantepec (state of Oaxaca, Mexico). Only non-myrmecophytes that grew at the same sites as myrmecophytes were selected, so that putative site effects on EFN secretion patterns could not cause differences among species (see Methods for a description of how EFN flow was induced and quantified). EFN secretion by all non-myrmecophytes (*Acacia cochliacantha*, *A. farnesiana*, *A. macracantha*, *A. pennatula*, *Leucaena leucocephala*, *Piptadenia flava* and *Prosopis juliflora*) increased after mechanically damaging leaves or after jasmonic acid application, and was five to more than ten times higher on treated than on control twigs (Fig. 1). In contrast, EFN secretion did not respond to leaf damage in any of the myrmecophytes (*Acacia chiapensis*, *A. collinsii*, *A. cornigera*, *A. globulifera* and *A. hindsii*, Fig. 1); all species secreted EFN constitutively at high rates. Therefore, the same trait was inducible in some, but constitutive in other species of the same genus, and there was an obvious relation to its different functions that matches the predictions of Agrawal & Rutter¹⁸. The induction of EFN secretion in non-myrmecophytes can guide non-specialized ants to plant parts that are currently under attack, whereas ants inhabiting myrmecophytes are provided with a permanent EFN flow at rates which in most cases are higher than those in induced non-myrmecophytes (Fig. 1).

EFN secreted by only one myrmecophyte (*A. collinsii*) responded to exogenous jasmonic acid application (Fig. 1), thus raising the question of whether the octadecanoid pathway is active in the species investigated here. Leaves of four myrmecophytes and four non-myrmecophytes were damaged mechanically in the field to

quantify endogenous jasmonic acid levels. Mechanical damage led to a significant increase (repeated measures analysis of variance (ANOVA): $P < 0.05$ for all eight species) in endogenous jasmonic acid in all species investigated (Fig. 2), demonstrating—along with the results of the inhibitor experiment (Fig. 3)—a functioning octadecanoid signalling cascade in all species. Retaining this pathway intact allows the plants to keep all other jasmonic-acid-responsive traits inducible (see refs 14, 15 for the high number of traits regulated through the octadecanoid pathway).

For further investigations of the role of jasmonic acid we used phenidone (1-phenyl-3-pyrazolidinone; Sigma-Aldrich), an inhibitor of lipoxygenases¹⁹. Endogenous jasmonic acid synthesis by *Acacia* leaves in response to mechanical damage is suppressed to a large degree (but not completely) when leaves are treated 30 min in advance and immediately after mechanically damaging them with a 2 mmol aqueous solution of phenidone (data not shown). The same treatment led to a suppression of the response in EFN secretion to damage in the non-myrmecophytes and to a strong decrease in EFN secretion by myrmecophytes (Fig. 3). EFN secretion (or the

response in EFN secretion to damage, respectively) could be restored by exogenous jasmonic acid application, showing that the suppression of EFN secretion by phenidone was indeed due to a suppressed endogenous jasmonic acid synthesis rather than caused by any side effects of phenidone.

Myrmecophytic *Acacia* species exhibit a constitutive EFN secretion; however, the octadecanoid pathway is active in these species. An increase in jasmonic acid, irrespectively of whether endogenous or exogenously applied, did not cause any increase in EFN secretion above the constitutive level in four of the myrmecophytes, even though EFN flow in myrmecophytes depends on jasmonic acid. It is obviously the responsiveness of EFN secretion to jasmonic acid, rather than an earlier part of the signalling cascade, which has been changed in these species. In myrmecophytes the basal levels of jasmonic acid (Fig. 2, levels at $t = 0$ min) are already sufficient to elicit maximum EFN secretion rates; however, lowering the jasmonic acid concentration still causes a decrease in EFN secretion.

How often has the regulation regime of EFN secretion changed during evolution, and which type is the ancestral one; an inducible or constitutive flow of EFN? So far only one study has attempted to reconstruct the phylogeny of an induced resistance. Thaler & Karban³ reported a repeated evolution of induced resistance to mites in cotton species. Unfortunately, this study suffered from the methodological difficulty of defining resistance as the number of mites developing on a given plant; this led to different results when the threshold separating 'resistant' from 'not resistant' was chosen differently.

The *Acacia* species investigated here belong to the monophyletic

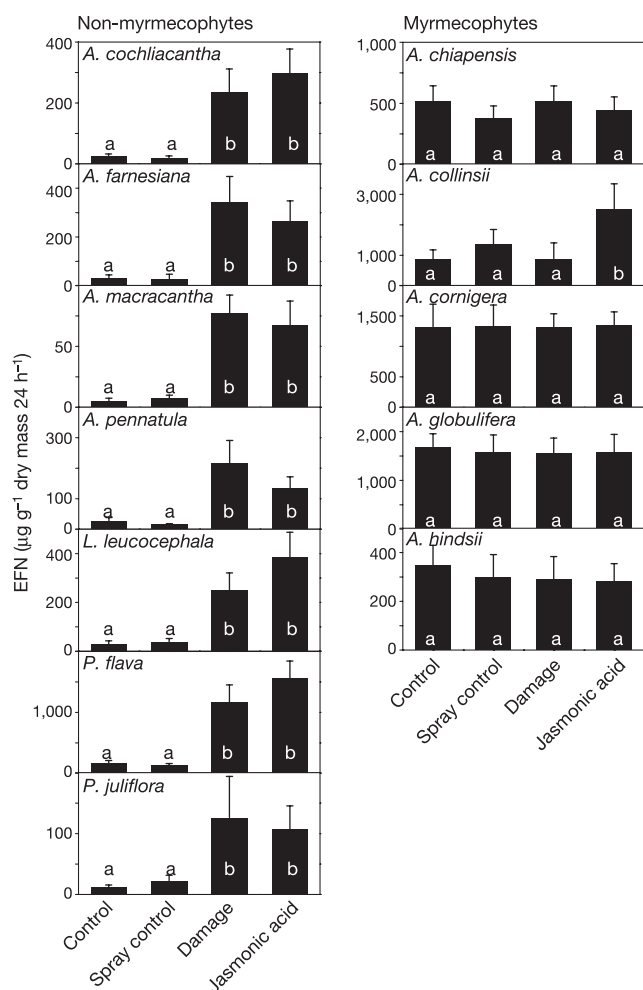


Figure 1 Extrafloral nectar secretion in response to different treatments. EFN secretion rates are given in μg soluble solids per g leaf dry mass per 24 h. Treatment (control, untreated control; spray control, water sprayed on leaves; damage, leaves mechanically damaged; jasmonic acid, aqueous solution of jasmonic acid sprayed on leaves) had a highly significant effect on EFN secretion by all non-myrmecophytes and *A. collinsii* (repeated measures ANOVA: $P < 0.01$). Different letters indicate significant differences among treatments ($P < 0.05$ according to least significant difference (LSD) post-hoc analysis). Sample size = 10 individual plants per species. Error bars indicate standard errors.

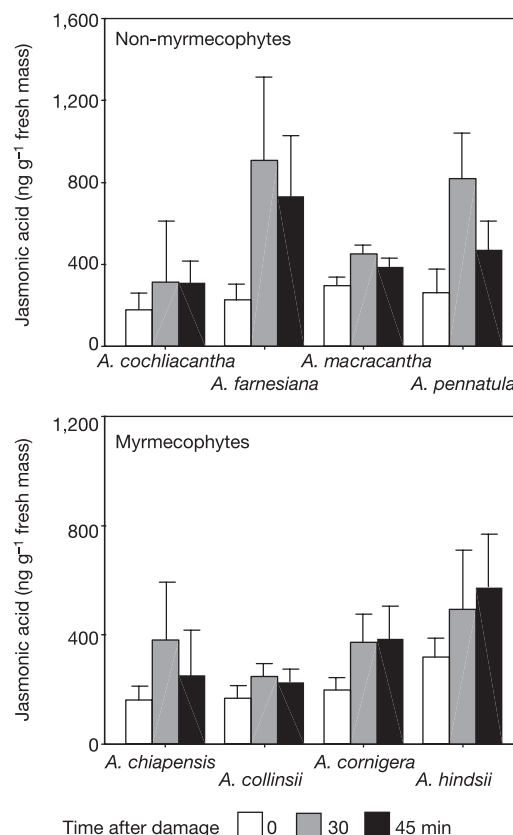


Figure 2 Endogenous concentrations of jasmonic acid. Endogenous jasmonic acid (ng per g fresh mass) in leaves of eight *Acacia* species at different times after mechanically damaging leaves. Sample size = 3 plant individuals per species. Error bars indicate standard errors.

subgenus *Acacia*, in which the neotropical representatives are embedded as a distinct clade^{20,21} (see Supplementary Materials). Inducible EFN secretion was observed in several species of subgenus *Acacia* as well as in non-*Acacia* species and has recently also been described for *Phaseolus lunatus*, a leguminous species belonging to the subfamily Faboideae⁶. Constitutive EFN flow was observed only in myrmecophytic *Acacia* species, indicating that this is the derived state. However, the myrmecophytic *Acacias* investigated are not included in published molecular phylogenetic surveys. To confirm their phylogenetic position, comparative sequencing of the *trnL-trnF* region—including the *trnL* intron and the spacer between the *trnL* and *trnF* genes—and of the *trnK* intron of the chloroplast DNA, was conducted for up to three individuals for the species studied (except *A. globulifera*) (see Methods). The alignment, using sequences for both loci of *Mimosa tenuiflora* (gi32365198 and gi27923094, ref. 21) as the outgroup, resulted in 3,302 aligned positions. Two hundred positions contained gaps indicating 28 insertions or deletions ('indels'). By eliminating length mutations in the polyA and polyT regions, five indels were potentially informative, only one of which is located within the species of *Acacia* studied. Employing indels as characters in phylogenetic analyses did not alter the tree topology. Maximum parsimony analysis yielded 57 most parsimonious trees with a length of 269 steps, a consistency index (CI) of 0.88 and a retention index (RI) of 0.83. The myrmecophytic *Acacia* taxa investigated are monophyletic within the subgenus *Acacia*, and, with *Mimosa* as the outgroup, *Piptadenia*, *Prosopis* and *Leucaena* are subsequent sisters to the *Acacia* species analysed (Fig. 4). This relationship is confirmed when sequences are analysed together with published data (Supplementary Figs). *A. farnesiana* (inducible EFN flow) is intercalating between a group containing inducible and constitutive species, indicating a paraphyly of inducible EFN secretion within

Acacia and thus supporting the assumption that this character is the ancestral state.

Additional support for the relationships, as observed based on chloroplast DNA sequences, was obtained by further analysis of genetic similarities using amplified fragment length polymorphism (AFLP) fingerprinting. AFLP markers are mainly composed of nuclear loci and thus reflect nuclear inheritance. Two hundred and eighteen highly polymorphic band levels were achieved by using four primer pairs. Different individuals of the same species clustered together in all cases, and the results of the cluster analysis corresponded in general to the maximum parsimony analysis of chloroplast loci (Fig. 4). The species exhibiting constitutive EFN flow form one cluster together with some species showing inducible EFN flow, whereas *A. farnesiana*, *P. flava*, *P. juliflora* and *L. leucocephala* (all inducible) form distinct clusters. Myrmecophytic *Acacia* species are positioned in two clusters and in this respect are not in congruence with the chloroplast DNA phylogeny.

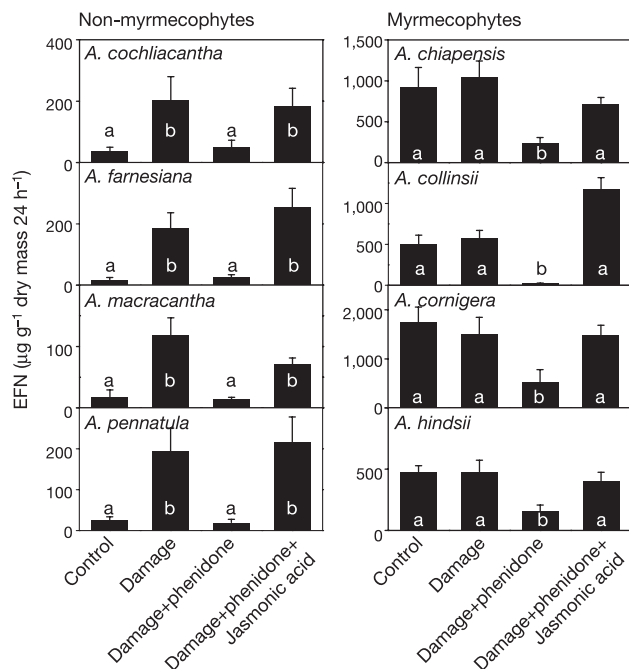
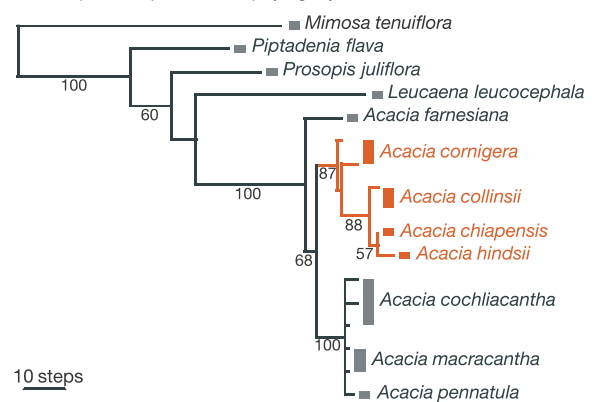


Figure 3 Extrafloral nectar secretion rates after inhibition of jasmonic acid synthesis. Endogenous jasmonic acid synthesis was inhibited by application of phenidone, and EFN secretion (μg soluble solids per g leaf dry mass per 24 h) was quantified one day later. Treatment significantly affected EFN secretion (repeated measures ANOVA: $P < 0.01$ for all species). Different letters indicate significant differences among treatments ($P < 0.05$ according to LSD post-hoc analysis). Sample size = 10 plant individuals per species. Error bars indicate standard errors.

a Chloroplast sequence data phylogeny



b AFLP cluster analysis

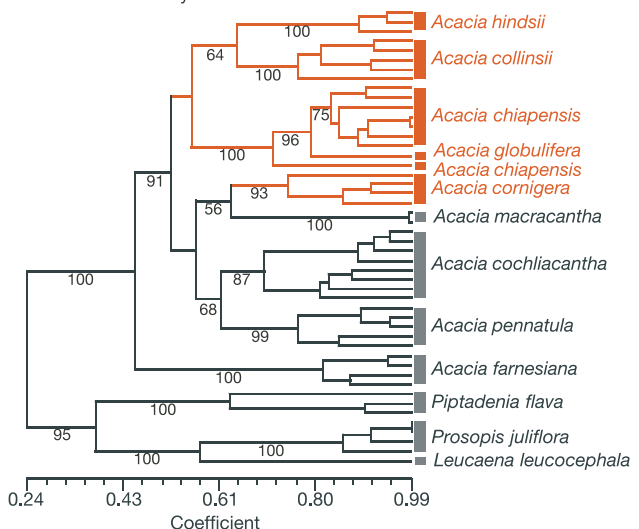


Figure 4 Phylogenetic relationships. **a**, **b**, Phylogenetic reconstruction based on a combined data matrix of the chloroplast DNA markers *trnL-trnF* region and *trnK* intron (**a**) (as a strict consensus tree from the 57 most parsimonious trees, shown as a phylogram), and genetic similarities among investigated species as revealed by cluster analysis of AFLP data (**b**) (as UPGMA from the Dice coefficient matrix). Myrmecophytic species (constitutive nectar secretion) are marked red. For both data sets several individuals have been analysed from most species. Equal sequences had been united into one taxonomical unit. Bootstrap values are given below the corresponding branch. Lengths of branches indicate the number of mutation steps (bar: ten steps) in the displayed phylogram.

A. cornigera appeared to be very similar to *A. macracantha* and was therefore nested within the non-mymecophytic Acacias. Finger-print analysis can be influenced by introgression of single loci. The high similarity of *A. cornigera* with *A. macracantha* could thus result from hybridization²²; however, this relationship is only weakly supported by bootstrap analysis.

In the AFLP fingerprints, the dissimilarity among species with induced EFN flow was higher than among species with constitutive EFN flow; however, both groups are comparable in numbers of both species and samples. This difference and the order of the corresponding samples in the cluster analysis confirmed the main result of chloroplast DNA sequencing; that is, the young age and derived character of constitutive EFN secretion.

For cotton, Thaler & Karban³ reported inducibility of resistance to be the derived state. In *Acacia*, it is obviously the constitutive EFN secretion that emerged as a derived trait during the evolution of myrmecophytism, showing a functional change in the way host plants guide defending animals while a facultative mutualism has been converted to an obligate one. Quantitative variation in the level of induced resistance traits have been reported, for example, for volatiles²³ or secondary metabolites²⁴. In contrast, our study shows qualitative differences: the inducible phenotype of a resistance trait has evolved to a constitutive one in response to modified functional demands. □

Methods

All the plants investigated were shrubs 1–2.5-m-tall, growing under field conditions. Species were determined following Janzen⁹ and by comparison with voucher specimens from the Herbarium MEXU, UNAM, Mexico City.

Induction and inhibition of EFN secretion

Ten plants per species were used in each of the two experiments on EFN induction. Four twigs per plant were matched for age, within-plant location and leaf number, and randomly subjected to different treatments. The five youngest fully expanded leaves of each twig were protected from nectar consumers by applying Tangletrap (Tanglefoot) around the twig base and putting the twigs into gauze bags. 'Control' leaves received no additional treatment. 'Damage' leaves were treated with a metal brush resulting in three to five holes per leaflet. On the 'jasmonic acid' leaves, an aqueous 1 mM solution of jasmonic acid was sprayed until there was runoff, and the same amount of water was sprayed on 'spray control' leaves. 'Damage + phenidone' leaves were sprayed two times with a 2 mmol aqueous solution of phenidone 30 min before and immediately after mechanically damaging them as described above, and 'damage + phenidone + jasmonic acid' received two additional sprays of jasmonic acid solution during the 30 min after the final phenidone application. EFN secretion was quantified in all cases 24 h after treatment as described previously^{7,25}. For quantification of endogenous jasmonic acid levels, leaves were damaged as in the 'damage' treatment and were harvested at $t = 0$ min and after 30 and 45 min. Extraction and quantification of endogenous jasmonic acid followed the protocol of Koch *et al.*²⁶. Shortly afterwards, 1.0 g of leaf tissue was frozen and [9,10-²H₂] dihydro-jasmonic acid was added as an internal standard²⁷. Jasmonic acid was extracted and purified using NH₂-propyl SPE cartridges (Varian). Detection and quantification of jasmonic acid was done by GC-SIM-MS (gas chromatography-selected ion monitoring-mass spectrometry) without further purification.

Phylogenetic analysis and genetic distances

Leaves were collected in the field and dried over silica gel at ambient temperature. DNA was extracted from roughly 40 mg of dry leaf material with DNeasy Plant Mini Kit (Qiagen). AFLP analysis was performed according to the protocol provided with the 'AFLP Core Reagent Kit' (Invitrogen) using the restriction enzymes *EcoRI* and *MseI*. Four primer pairs differing by their selective nucleotides (*E*_{AAC} versus *M*_{CTC}, *M*_{CAT} and *M*_{CTA}, and *E*_{AAG} versus *M*_{CAT}) were used for the final polymerase chain reaction (PCR) amplification. The PCR products were analysed in 5% Urea-PAGE. AFLP patterns were revealed by exposure to X-ray films (LifeRay, Ferrania). Band levels were scored as present (1) or absent (0). Estimation of genetic similarities and cluster analysis were performed with the NTSys pc Version 2.10 (Applied Biostatistics) using the Dice coefficient (corresponding to Nei-Li distances) for the analysis of banding patterns (Manual for NTSys pc). Cluster analysis was performed as unweighted pair group method using arithmetic averages (UPGMA).

For analysing chloroplast DNA sequences, amplification was performed as described previously²⁸. The entire *trnL-trnF* region was amplified using the primers C and F (ref. 29), and the *trnK* intron was amplified in two parts using the primers 2-trnK-3914F and 16-trnK-2R (ref. 30) and the internal primers AC1100F (GCCGTCAGTGTGGAAATTCC) and AC1300R (CAGTATCGAAGGGTTGAATC), designed after comparison with published sequences for *Acacia*²¹. Sequencing was performed using an ABI 377 (Applied Biosystems) according to the manufacturer's protocol, applying the same primers as for amplification. For the *trnK* intron AC1700F (GGAAATCCATTCTGGCTTC) was used as an additional sequencing primer. The sequences were aligned manually and maximum parsimony analysis was performed using PAUP 4.0b10 (ref. 28) for the single data sets and

a combined matrix from both loci. In addition, maximum likelihood analysis and neighbour joining with PAUP were performed and revealed analogous results.

Received 5 April; accepted 21 May 2004; doi:10.1038/nature02703.

1. Walling, L. L. The myriad plant responses to herbivores. *J. Plant Growth Reg.* **19**, 195–216 (2000).
2. Karban, R. & Baldwin, I. T. *Induced Responses to Herbivory* (University of Chicago Press, Chicago and London, 1997).
3. Thaler, J. S. & Karban, R. A phylogenetic reconstruction of constitutive and induced resistance in *Gossypium*. *Am. Nat.* **149**, 1139–1146 (1997).
4. Price, P. W. *et al.* Interactions among three trophic levels: influence of plants on interactions between insect herbivores and natural enemies. *Annu. Rev. Ecol. Syst.* **11**, 41–65 (1980).
5. Heil, M. & McKey, D. Protective ant-plant interactions as model systems in ecological and evolutionary research. *Annu. Rev. Ecol. Syst.* **34**, 425–453 (2003).
6. Heil, M. Induction of two indirect defences benefits Lima bean (*Phaseolus lunatus*, Fabaceae) in nature. *J. Ecol.* **92**, 527–536 (2004).
7. Heil, M. *et al.* Extrafloral nectar production of the ant-associated plant, *Macaranga tanarius*, is an induced, indirect, defensive response elicited by jasmonic acid. *Proc. Natl Acad. Sci. USA* **98**, 1083–1088 (2001).
8. Wäckers, E. L., Zuber, D., Wunderlin, R. & Keller, F. The effect of herbivory on temporal and spatial dynamics of foliar nectar production in cotton and castor. *Ann. Bot.* **87**, 365–370 (2001).
9. Janzen, D. H. *Swollen-Thorn Acacias of Central America* (Smithsonian Institution Press, Washington DC, 1974).
10. Agrawal, A. A. Induced responses to herbivory and increased plant performance. *Science* **279**, 1201–1202 (1998).
11. Baldwin, I. T. Jasmonate-induced responses are costly but benefit plants under attack in native populations. *Proc. Natl Acad. Sci. USA* **95**, 8113–8118 (1998).
12. Thaler, J. S. Jasmonate-inducible plant defences cause increased parasitism of herbivores. *Nature* **399**, 686–688 (1999).
13. Kessler, A. & Baldwin, I. T. Defensive function of herbivore-induced plant volatile emissions in nature. *Science* **291**, 2141–2144 (2001).
14. Creelman, R. A. & Mullet, J. E. Biosynthesis and action of jasmonates in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **48**, 355–381 (1997).
15. Wasternack, C. & Parthier, B. Jasmonate-signalled plant gene expression. *Trends Plant Sci.* **2**, 302–307 (1997).
16. Heil, M., Baumann, B., Krüger, R. & Linsenmair, K. E. Main nutrient compounds in food bodies of Mexican *Acacia* ant-plants. *Chemecology* **14**, 45–52 (2004).
17. Raine, N. E., Willmer, P. & Stone, G. N. Spatial structuring and floral avoidance behavior prevent ant-pollinator conflict in a Mexican ant-acacia. *Ecology* **83**, 3086–3096 (2002).
18. Agrawal, A. A. & Rutter, M. T. Dynamic anti-herbivore defense in ant-plants: the role of induced responses. *Oikos* **83**, 227–236 (1998).
19. Cucurou, C., Battioni, J. P., Thang, D. C., Nam, N. H. & Mansuy, D. Mechanisms of inactivation of lipoxygenases by phenidone and BW755C. *Biochemistry* **30**, 8964–8970 (1991).
20. Miller, J. T. & Bayer, R. J. Molecular phylogenetics of *Acacia* subgenera *Acacia* and *Aculeiferum* (Fabaceae: Mimosoideae), based on the chloroplast *matK* coding sequence and flanking *trnK* intron spacer regions. *Aust. Syst. Bot.* **16**, 27–33 (2003).
21. Miller, J. T., Grimes, J. W., Murphy, D. J., Bayer, R. J. & Ladiges, P. Y. A phylogenetic analysis of the Acaciae and Ingeae (Mimosoideae: Fabaceae) based on *trnK*, *matK*, *psbA-trnH*, and *trnL/trnF* sequence data. *Syst. Bot.* **28**, 558–566 (2003).
22. Wolfe, A. D., Xiang, Q. Y. & Kephart, S. R. Diploid hybrid speciation in *Penstemon* (Scrophulariaceae). *Proc. Natl Acad. Sci. USA* **95**, 5112–5115 (1998).
23. Loughrin, J. H., Manukian, A., Heath, R. R. & Tumlinson, J. H. Volatiles emitted by different cotton varieties damaged by feeding beet armyworm larvae. *J. Chem. Ecol.* **21**, 1217–1227 (1995).
24. Siemens, D. H. & Mitchell-Olds, T. Evolution of pest-induced defenses in *Brassica* plants: Tests of the theory. *Ecology* **79**, 632–646 (1998).
25. Heil, M., Fiala, B., Baumann, B. & Linsenmair, K. E. Temporal, spatial and biotic variations in extrafloral nectar secretion by *Macaranga tanarius*. *Funct. Ecol.* **14**, 749–757 (2000).
26. Koch, T., Krumm, T., Jung, V., Engelberth, J. & Boland, W. Differential induction of plant volatile biosynthesis in the lima bean by early and late intermediates of the octadecanoid-signaling pathway. *Plant Physiol.* **121**, 153–162 (1999).
27. Baldwin, I. T. *et al.* Quantification, correlations and manipulation of wound-induced changes in jasmonic acid and nicotine in *Nicotiana glauca*. *Planta* **201**, 397–404 (1997).
28. Meimberg, H., Wistuba, A., Dittrich, P. & Heubl, G. Molecular phylogeny of Nepenthaceae based on cladistic analysis of plastid *trnK* intron sequence data. *Plant Biol.* **3**, 154–175 (2001).
29. Taberlet, P., Gielly, L., Pautou, G. & Bouvet, J. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Mol. Biol.* **17**, 1105–1109 (1991).
30. Johnson, L. A. & Soltis, D. E. *MatK* DNA sequences and phylogenetic reconstruction in Saxifragaceae s. str. *Syst. Bot.* **19**, 143–156 (1994).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank D. Davidson, J. Gershenzon, I. Baldwin, G. Stone, C. Kost and G. Pohnert for critically reading earlier versions of the manuscript, K. Heil and B. Baumann for help with field work, A. Kreiner and F. Thiele for taking care of greenhouse plant, and M. Sousa for species determination. Financial support by the Deutsche Forschungsgesellschaft and the Max-Planck-Society is acknowledged.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.H. (Heil_Martin@web.de). The *trnK* intron and the *trnL-trnF* region sequences are published in GenBank under accession numbers AY574092–AY574105 (*trnK* intron) and AY574106–AY574114 (*trnL-trnF* region).

Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia

K. S. Li¹, Y. Guan^{1,2}, J. Wang^{1,2}, G. J. D. Smith^{1,2}, K. M. Xu^{1,2}, L. Duan^{1,2}, A. P. Rahardjo³, P. Puthavathana⁴, C. Buranathai⁵, T. D. Nguyen⁶, A. T. S. Estoepongastie³, A. Chaisingh⁵, P. Auewarakul⁴, H. T. Long⁷, N. T. H. Hanh⁷, R. J. Webby⁸, L. L. M. Poon², H. Chen^{1,2}, K. F. Shortridge^{1,2}, K. Y. Yuen², R. G. Webster^{2,8} & J. S. M. Peiris^{1,2}

¹Joint Influenza Research Centre (SUMC & HKU), Shantou University Medical College, Shantou, Guangdong 515031, China

²Department of Microbiology, The University of Hong Kong, Queen Mary Hospital, Hong Kong SAR, China

³Fakultas Kedokteran Hewan, Universitas Airlangga, Surabaya 60115, Indonesia

⁴Department of Microbiology, Sriraj Hospital, Bangkok 10700, Thailand

⁵Department of Livestock Development, National Institute of Animal Health, Bangkok 10900, Thailand

⁶Department of Virology, National Institute of Veterinary Research, Ministry of Agriculture and Rural Development, Hanoi, Vietnam

⁷Virology Department, National Institute of Hygiene and Epidemiology, Hanoi, Vietnam

⁸Virology Division, Department of Infectious Diseases, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA

A highly pathogenic avian influenza virus, H5N1, caused disease outbreaks in poultry in China and seven other east Asian countries between late 2003 and early 2004; the same virus was fatal to humans in Thailand and Vietnam¹. Here we demonstrate a series of genetic reassortment events traceable to the precursor of the H5N1 viruses that caused the initial human outbreak in Hong Kong in 1997 (refs 2–4) and subsequent avian outbreaks in 2001 and 2002 (refs 5, 6). These events gave rise to a dominant H5N1 genotype (Z) in chickens and ducks that was responsible for the regional outbreak in 2003–04. Our findings indicate that domestic ducks in southern China had a central role in the generation and maintenance of this virus, and that wild birds may have contributed to the increasingly wide spread of the virus in Asia. Our results suggest that H5N1 viruses with pandemic potential have become endemic in the region and are not easily eradicable. These developments pose a threat to public and veterinary health in the region and potentially the world, and suggest that long-term control measures are required.

The Asian outbreak of highly pathogenic avian influenza H5N1 disease in poultry in 2003 and 2004 was unprecedented in its geographical extent, and its transmission to humans was an ominous sign¹. To trace the ecological and genetic origins of these outbreaks, we compared H5N1 viruses recently isolated from poultry in Indonesia, Thailand and Vietnam as well as from humans in Thailand and Vietnam with 253 H5N1 isolates obtained during prospective surveillance of live poultry markets in Hong Kong and in Guangdong, Hunan and Yunnan provinces, China, from 2000 to 2004 (Fig. 1). Results of this surveillance are summarized in Table 1. Since 2001, H5N1 viruses have continued to circulate in mainland China with a seasonal pattern, peaking from October to March, when the mean temperature is below 20 °C (Fig. 2). The survival and viability of influenza A virus are known to increase at lower environmental temperatures⁷. H5N1 viruses were isolated exclusively from aquatic poultry during 2000; however, from 2001 onwards they were isolated from both aquatic and terrestrial poultry, although the rate of isolation remained greatest in ducks.

The genes of the virus isolates that encode the surface antigens haemagglutinin (HA) and neuraminidase (NA) were derived from the Goose/Guangdong/1/96 (Gs/Gd)-like lineage. Six genes, which

encode internal viral proteins, arose from many other sources through reassortment and served as the basis for assignment to different genotypes (Fig. 3). By 2001, six H5N1 reassortants (genotypes A, B, C, D, E and X₀) had been isolated from aquatic and, for the first time since 1997, terrestrial poultry^{4,5}. From 2002 onwards, eight new H5N1 genotypes (V, W, X₁, X₂, X₃, Y, Z and Z⁺) were detected. Genotypes A, C, D and E and their common precursor Gs/Gd were no longer found, suggesting that later genotypes had acquired a survival advantage by means of adaptation. At least nine genotypes of H5N1 viruses continued to circulate in southern China in 2002. All genotypes, except for Gs/Gd and X₀–X₃, had a five-amino-acid deletion (position 80–84) in the NS1 protein. Similarly, viruses isolated in 2002 and later, except for genotypes B, W and Z⁺, had a 20-amino-acid deletion in the stalk of the NA molecule (position 49–68). Deletion in the NA stalk may be associated with adaptation of influenza viruses to land-based poultry⁸.

Since January 2002, genotype Z, which contains both the NA and NS1 deletions, has become the dominant H5N1 virus in southern China (Table 1 and Fig. 3). In February 2003, human H5N1 disease was diagnosed for the first time since December 1997. The human isolates (A/HK/212/03 and A/HK/213/03) had the same gene constellation as genotype Z but lacked the NA stalk deletion, and were designated genotype Z⁺ (Fig. 3)⁶. Sixty-two H5N1 virus isolates from 2003 were genetically sequenced and 60 of them belonged to genotype Z (Table 1). All of the viruses that caused outbreaks in Indonesia, Thailand and Vietnam in late 2003 and early 2004 (refs 1, 9) were genotype Z viruses (Fig. 4).

Although all viruses tested derived their HA genes from Gs/Gd-like viruses, the Z and Z⁺ genotypes, as well as the single known isolate of genotype V (Ck/ST/4231/03), formed a distinct sublineage (2002/04-V, Z, Z⁺; 81% bootstrap support) that included viruses from Hong Kong; Guangdong, Hunan and Yunnan provinces of mainland China; Indonesia; Thailand; and Vietnam (Fig. 4a). Within this sublineage, viruses isolated from humans and poultry in Thailand and Vietnam formed a separate group (96% bootstrap) most closely related to H5N1 viruses isolated from poultry and wild birds in Hong Kong (genotypes Z and Z⁺). Viruses isolated in Indonesia formed a separate group related to those isolated in Yunnan (Fig. 4a).

Phylogenetic relationships of the other seven gene segments were broadly consistent with those of the HA gene (Fig. 4b; see also Supplementary Figs 1 and 2). The M and NS genes were closest to those of influenza viruses of diverse subtypes isolated from ducks in southern China, suggesting that aquatic avian viruses were the gene



Figure 1 Map of China showing Hong Kong and Guangdong, Hunan and Yunnan provinces, where influenza surveillance was conducted.

donors for these new reassortants (Fig. 4b; see also Supplementary Fig. 2). Bi-directional transmission of influenza virus between terrestrial and aquatic poultry gave rise to new H9N2 reassortants¹⁰, and a similar mechanism may generate novel H5N1 reassortants.

It is notable that in the short time since its emergence in 2002, genotype Z has replaced genotypes A–E, X and Y to become dominant in both aquatic and terrestrial poultry in this region (Table 1). To define the genetic stability of this new gene constellation, we analysed the rates of non-synonymous (K_a) and synonymous (K_s) nucleotide substitutions in six internal gene segments of genotype Z viruses isolated in 2002–04. A K_a/K_s ratio >1 suggests evidence of positive natural selection¹¹. Of these internal genes, the M2 gene was under positive selection pressure in late 2002 to early 2003 but under less selection pressure in late 2003 to early 2004. The NS1 and NS2 genes, acquired in late 2000, were also under positive selection pressure (Supplementary Table 1). These findings suggest that the gene constellation of genotype Z viruses has not yet fully adapted to poultry, and this raises the possibility that they may continue to evolve through mutation or reassortment to achieve greater viral fitness¹².

The presence of residue Asp 31 in the M2 protein invariably confers resistance to the amantadines¹³, a group of antiviral drugs used for treatment of human influenza. Sequence analysis revealed that Asp 31 was present in all avian and human genotype Z viruses isolated in Thailand and Vietnam, but in only one of six viruses isolated from poultry in Indonesia. Asp 31 was also observed in some genotype B, Y and Z⁺ viruses. The distribution of this mutation in different virus genotypes suggests that it was independently acquired rather than having descended from a single lineage of amantadine-resistant M2 genes (Fig. 4b).

Mutations Ser64Ala and Glu66Ala in the M2 protein were also observed in genotype Z and Z⁺ viruses from highly pathogenic avian influenza H5N1 outbreaks in Kowloon and Penfold parks in Hong Kong in December 2002, from viruses isolated from humans in Hong Kong in February 2003, and from viruses isolated in Thailand and Vietnam in 2003–04. The remaining genotype Z viruses, including those isolated from Indonesia and mainland China in 2003–04, maintained Ser 64 in the M2 protein. Residue Ser 64 is the predominant site of post-translational phosphorylation in the M2 protein^{14,15}. The mutation Ser64Ala induces no significant change in ion-channel activity¹⁴, and phosphorylation of the cytoplasmic tail does not affect intracellular transport of M2 protein or viral assembly¹⁵. Therefore, the biological role of phosphorylation at position 64 of M2 is still unclear. The biological significance of residue 66 of the M2 ion channel has not been reported.

The molecular determinants of H5N1 transmission to humans in Vietnam and Thailand in 2004 are unclear. In H5N1 viruses recently isolated from humans and poultry in Thailand and Vietnam, amino

acid residues at the receptor-binding pocket of HA1—that is, positions Gln 222 and Gly 224 (positions 226 and 228 for H3 influenza numbering)—retain configurations (2,3-NeuAcGal linkages) predicted to have affinity for avian cell-surface receptors¹⁶. The substitution Ser227Asn, identified in viruses isolated from two patients with H5N1 influenza after visiting Fujian Province, China, in 2003 (ref. 6), was not seen in any other H5N1 viruses. Other amino acid residues relevant to receptor binding (residues 91, 130–134, 149, 151, 179, 186, 190–191, 220–225) were identical to

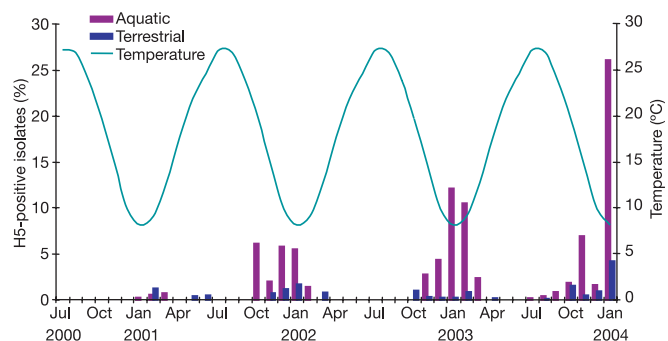


Figure 2 Seasonality of the isolation of avian H5N1 viruses from domestic poultry in mainland China during July 2000 to January 2004 (see Table 1). The mean monthly temperature in southern China (approximated from the monthly average temperatures of the cities Changsha, Kunming and Xiamen) is shown for reference.

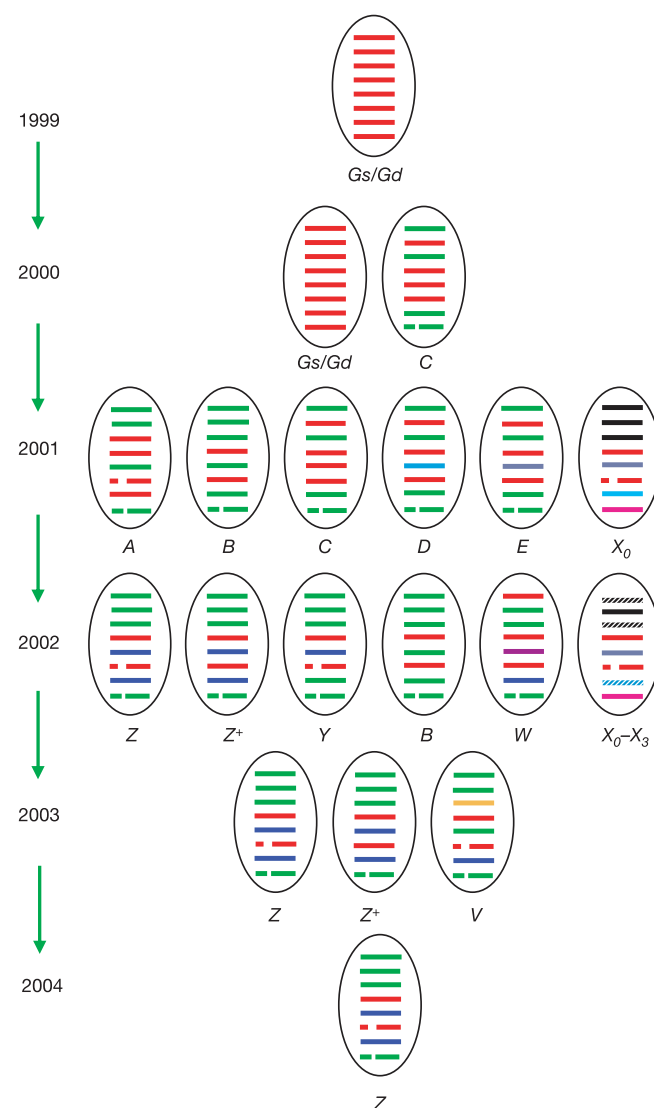


Figure 3 The genotypes of H5N1 influenza virus reassortants from eastern Asia. The eight gene segments are (horizontal bars starting at the top downwards): PB2, PB1, PA, HA, NP, NA, M and NS. Each colour represents a virus lineage (red indicates origin from Gs/Gd/1/96). Genotypes (indicated by letters) were defined by gene phylogeny: a distinct phylogenetic lineage with bootstrap support $\geq 70\%$ ($\geq 50\%$ for M, NP and PA genes) indicated a common origin. Genotypes A, B and C were reassortants of Gs/Gd/1/96 and one or more aquatic avian viruses. Genotype D was created when the NP gene of genotype C was replaced by that of a Dk/HK/Y280/97-like virus (H9N2 subtype). Genotype E was created when the NP gene of genotype C was replaced by that of another avian virus. Further reassortment of genotype E with other aquatic avian influenza viruses gave rise to the genotypes X₀–X₃, distinguished by the sources of their PB2, PA and NS genes. Genotype W differs from genotype B only in its PB2, NP and M genes. Further reassortment of genotype A or B with other aquatic avian viruses gave rise to genotypes V, Y, Z and Z⁺. Alternatively, genotype V may have resulted from reassortment of genotype Z with other aquatic avian viruses.

those of A/HK/156/97 and Gs/Gd-like viruses³, with the exception of A/Vietnam/3046/04, which had an Ala134Val mutation.

The HA molecules of most genotype Z viruses isolated since late 2002 in Hong Kong, of two out of six isolates from Indonesia, and of all isolates from Thailand, Vietnam and Yunnan Province in late 2003 and early 2004 had acquired a potential N-linked glycosylation site at positions 154–156. Glycosylation at this site, adjacent to the receptor-binding³ and antigenic sites¹⁷ at the globular tip of the H5 influenza HA molecule (Supplementary Fig. 3), is capable of altering the receptor-binding profile¹⁸ and may help the virus to evade the host antibody response.

Lys 627 in the PB2 protein has been associated with increased virulence of H5N1 viruses in mice¹⁹ and of H7N7 viruses in humans²⁰. Three out of four H5N1 human virus isolates from Vietnam had this mutation, but the human virus from Thailand did not, nor did any other avian influenza viruses tested. No recent H5N1 viruses characterized in this study had Glu 92 in the NS1 protein, which is reportedly associated with increased virulence in pigs²¹.

The apparently simultaneous occurrence of H5N1 outbreaks across eastern Asia remains unexplained, but the presence of H5N1 viruses in dead migratory birds suggests that wild bird populations may be involved. In Hong Kong between late 2002 and the time of this report, genotype Z⁺ H5N1 virus was isolated from a dead little egret (*Egretta garzetta*), and genotype Z viruses were isolated from two dead grey herons (*Ardea cinerea*), a black-headed gull (*Larus ridibundus*), a feral pigeon (*Columba livia*), a tree sparrow (*Passer montanus*) and a peregrine falcon (*Falco peregrinus*). In the gene phylogenies, the H5N1 viruses isolated from wild birds have either an out-group or sister-group relation to recent Thailand and Vietnam H5N1 isolates (Fig. 4; see also Supplementary Figs 1 and 2). The timing and distribution of the H5N1 infection in poultry in China from 2001 onwards (Fig. 2) coincides with the general period of winter bird migration to southern China; however, it is not known whether the H5N1 virus has become established in wild bird populations. The potential role of wild birds in the maintenance and spread of H5N1 viruses must be considered in strategies for regional control.

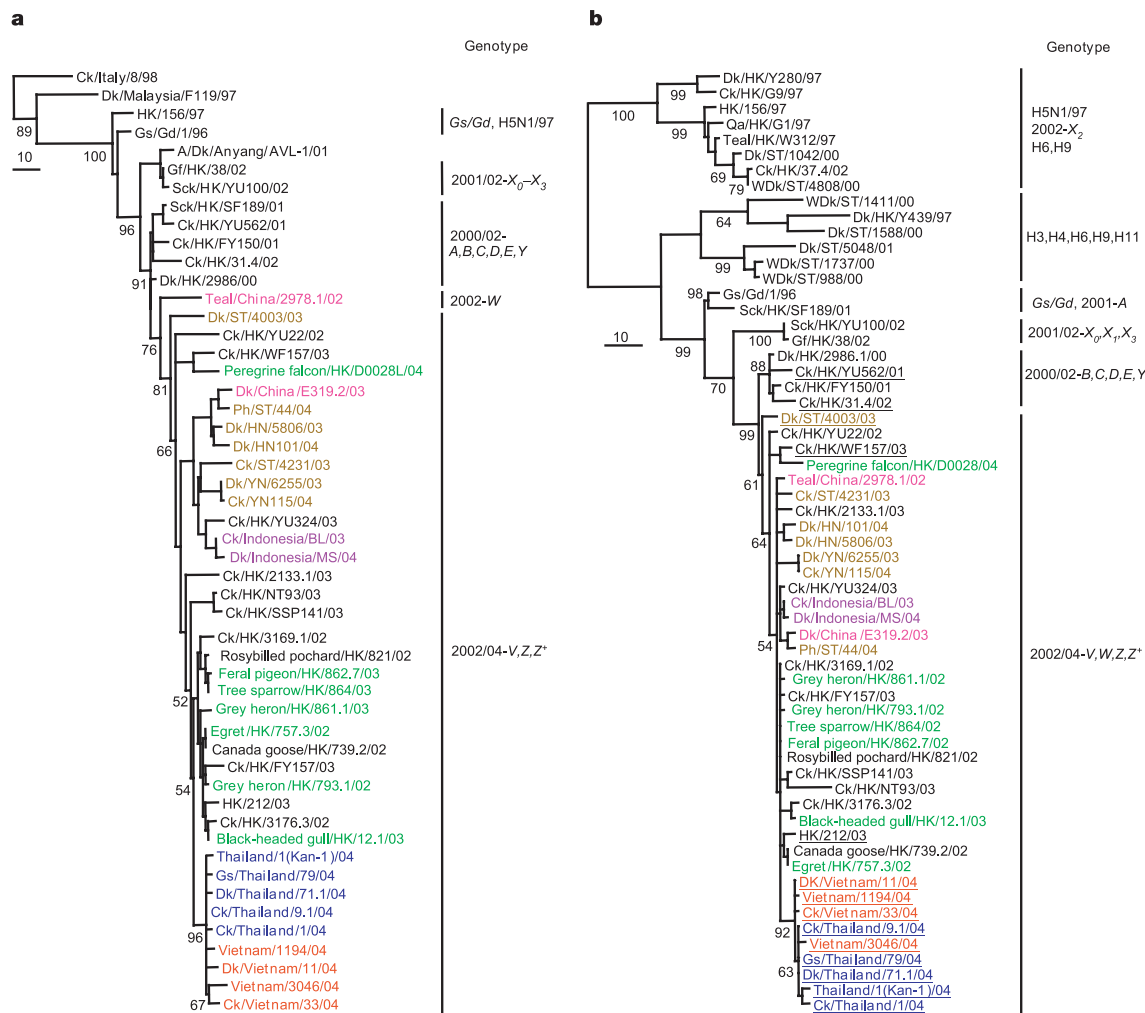


Figure 4 Phylogenetic relationships of the haemagglutinin (**a**) and matrix protein (**b**) genes of representative influenza A viruses isolated in southeastern Asia, including 2 of 6 from Indonesia, 5 of 8 from Thailand and 4 of 12 from Vietnam. Trees were generated by using maximum parsimony in the PAUP* program²⁵ (neighbour-joining analysis with the Tamura–Nei γ -model, implemented in the MEGA program²⁶, revealed the same relationships). Numbers below branches indicate bootstrap values from 1,000 replicates. Only bootstrap values that define important groups have been included owing to space constraints. Analysis was based on nucleotides 1–1012 (1,012 bp) of the HA gene and

90–945 (856 bp) of the M gene. The HA tree was rooted to A/tern/South Africa/61 and the M tree to A/equine/Prague/1/56. Scale bar, 10 nucleotide changes. Green text indicates viruses isolated from wild birds in Hong Kong; pink text indicates viruses from smuggled birds in China; and other colours show the country of origin of isolates from the late 2003 to early 2004 H5N1 outbreak. Underlined viruses have the amantadine-resistance mutation (Ser31Asn) in the M2 ion channel. Ck, chicken; Dk, duck; Gd, Guangdong; Gf, Guinea fowl; Gs, goose; HK, Hong Kong; HN, Hunan; Qa, quail; Sck, silky chicken; ST, Shantou; WDK, wild duck; YN, Yunnan.

Table 1 Analysis of H5N1 influenza viruses isolated from poultry in Hong Kong and mainland China during 2000–04

	2000		2001		2002		2003		2004	
	Aquatic	Terrestrial	Aquatic	Terrestrial	Aquatic	Terrestrial	Aquatic	Terrestrial	Aquatic	Terrestrial
Sampling										
Hong Kong										
H5N1-positive/total tested	33/533	0/8,256	37/606	36/16,116	14/578	323/14,691	4/3,694	112/12,146	1/756	0/1,807
Non-H5N1 isolates	16	715	11	983	10	369	12	288	0	20
Mainland China										
H5N1-positive/ total tested	0/445	0/1,891	38/2,579	10/3,197	73/4,539	8/4,059	297/7,209	50/10,308	152/1152	26/1,673
Non-H5N1 isolates	122	143	468	290	507	297	301	361	22	75
Total number sampled	11,125		22,498		23,867		33,357		5,388	
Genetic analysis										
Hong Kong										
Number of viruses analysed (genotypes detected)	11 (9Gs/Gd; 2C)	–	13 (4B, 9C)	24 (13A, 4B, 4C, 1D, 2E)	12 (1B, 8Z, 3Z ⁺)	100 (7B, 8X ₀ , 2X ₁ , 3X ₂ , 1X ₃ , 10Y, 69Z)	4 (4Z)	38 (37Z, 1Z ⁺)	1 (1Z)	–
Mainland China										
Number of viruses analysed (genotypes detected)	–	–	7 (2B, 5X ₀)	1 (1X ₀)	10 (3B, 5Z, 2W)	7 (7Z)	11 (11Z)	9 (8Z, 1V)	3 (3Z)	2 (2Z)
Number of genotypes	2		6		9		3		1	

Faecal droppings from apparently healthy poultry in live poultry markets in Hong Kong (2000–04), Guangdong (2000–04) and Hunan and Yunnan (2002–04) provinces were sampled monthly for influenza virus isolation. For each month that H5N1 virus was identified, one isolate was selected from each type of infected poultry for sequencing. During H5N1 disease outbreaks, additional isolates were sequenced. Of a total of 96,235 samples, 253 H5N1 virus isolates were genetically sequenced and analysed.

H5N1 virus is now endemic in poultry in Asia (Table 1) and has gained an entrenched ecological niche from which to present a long-term pandemic threat to humans. At present, these viruses are poorly transmitted from poultry to humans, and there is no conclusive evidence of human-to-human transmission. However, continued, extensive exposure of the human population to H5N1 viruses increases the likelihood that the viruses will acquire the necessary characteristics for efficient human-to-human transmission through genetic mutation or reassortment with a prevailing human influenza A virus. Furthermore, contemporary human H3N2 influenza viruses are now endemic in pigs in southern China²² and can reassort with avian H5N1 viruses in this 'intermediate host'. Therefore, it is imperative that outbreaks of H5N1 disease in poultry in Asia are rapidly and sustainably controlled. The seasonality of the disease in poultry, together with the control measures already implemented, are likely to reduce temporarily the frequency of H5N1 influenza outbreaks and the probability of human infection. However, complacency would be unwise. Governments in the region face an endemic and recurrent problem that presents a serious threat to human health. Although other countries in the region have been affected, Hong Kong has remained remarkably free of H5N1 outbreaks in poultry in 2004, thanks to preventive measures implemented over the past few years²³. □

Methods

Surveillance, virus isolation and characterization

Faecal droppings from apparently healthy poultry in live poultry markets in Hong Kong and in Guangdong, Hunan and Yunnan provinces were sampled monthly for influenza virus isolation. Methods used for virus isolation and characterization have been previously described²⁴. In addition to systematic surveillance in the poultry markets, sick or dead poultry from markets and farms in Hong Kong were similarly studied. This analysis included representative H5N1 viruses isolated from birds in Vietnam (*n* = 8), Thailand (*n* = 7) and Indonesia (*n* = 6), and from humans in Vietnam (*n* = 4) and Thailand (*n* = 1) during the 2003–04 H5N1 outbreak.

Genetic analysis

All sequences were edited with the Staden software package and aligned with ClustalX. Phylogenetic trees were generated using PAUP* version 4.0 (ref. 25) and MEGA version 2.1 (ref. 26). *K_a/K_s* analysis was conducted by the Pamilo–Bianchi–Li method as implemented in MEGA.

Received 13 April; accepted 14 June 2004; doi:10.1038/nature02746.

1. World Health Organization, Avian influenza A (H5N1). *Weekly Epidemiol. Rev.* **79**, 65–70 (2004).
2. Xu, X., Subbarao, K., Cox, N. J. & Guo, Y. Genetic characterization of the pathogenic influenza A/Goose/Guangdong/1/96 (H5N1) virus: similarity of its haemagglutinin gene to those of H5N1

viruses from the 1997 outbreaks in Hong Kong. *Virology* **261**, 15–19 (1999).
3. Claas, E. C. J. *et al.* Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus. *Lancet* **351**, 472–477 (1998).
4. Guan, Y., Shortridge, K. F., Krauss, S. & Webster, R. G. Molecular characterization of H9N2 influenza viruses: were they the donors of the “internal” genes of H5N1 viruses in Hong Kong? *Proc. Natl Acad. Sci. USA* **96**, 9363–9367 (1999).
5. Guan, Y. *et al.* Emergence of multiple genotypes of H5N1 avian influenza viruses in Hong Kong SAR. *Proc. Natl Acad. Sci. USA* **99**, 8950–8955 (2002).
6. Guan, Y. *et al.* H5N1 Influenza: A protean pandemic threat. *Proc. Natl Acad. Sci. USA* **101**, 8156–8161 (2004).
7. Shortridge, K. F. *et al.* Characterization of avian H5N1 influenza viruses from poultry in Hong Kong. *Virology* **252**, 331–342 (1998).
8. Matrosovich, M., Zhou, N. N., Kawaoka, Y. & Webster, R. G. The surface glycoproteins of H5 influenza viruses isolated from humans, chickens, and wild aquatic birds have distinguishable properties. *J. Virol.* **73**, 1146–1155 (1999).
9. Hien, T. T. *et al.* Avian influenza A (H5N1) in 10 patients in Vietnam. *N. Engl. J. Med.* **350**, 1179–1188 (2004).
10. Li, K. S. *et al.* Characterization of H9 subtype influenza viruses from the ducks of southern China: a candidate for the next influenza pandemic in humans? *J. Virol.* **77**, 6988–6994 (2003).
11. Presgraves, D. C., Balagopalan, L., Abmayr, S. M. & Orr, H. A. Adaptive evolution drives divergence of a hybrid inviability gene between two species of *Drosophila*. *Nature* **423**, 715–719 (2003).
12. Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M. & Kawaoka, Y. Evolution and ecology of influenza A viruses. *Microbiol. Rev.* **56**, 152–179 (1992).
13. Scholtissek, C., Quack, G., Klenk, H. D. & Webster, R. G. How to overcome resistance of influenza A viruses against adamantane derivatives. *Antiviral Res.* **37**, 83–95 (1998).
14. Holsinger, L. J., Shaughnessy, M. A., Micko, A., Pinto, L. H. & Lamb, R. A. Analysis of the posttranslational modifications of the influenza virus M2 protein. *J. Virol.* **69**, 1219–1225 (1995).
15. Thomas, J. M., Stevens, M. P., Percy, N. & Barclay, W. S. Phosphorylation of the M2 protein of influenza A virus is not essential for virus viability. *Virology* **252**, 54–64 (1998).
16. Ha, Y., Stevens, D. J., Skehel, J. J. & Wiley, D. C. X-ray structures of H5 avian and H9 swine influenza virus hemagglutinins bound to avian and human receptor analogs. *Proc. Natl Acad. Sci. USA* **98**, 11181–11186 (2001).
17. Kaverin, N. V. *et al.* Structure of antigenic sites on the haemagglutinin molecule of H5 avian influenza virus and phenotypic variation of escape mutants. *J. Gen. Virol.* **83**, 2497–2505 (2002).
18. Iwatsuki-Horimoto, K., Kanazawa, R., Sugii, S., Kawaoka, Y. & Horimoto, T. The index influenza A virus subtype H5N1 isolated from a human in 1997 differs in its receptor-binding properties from a virulent avian influenza virus. *J. Gen. Virol.* **85**, 1001–1005 (2004).
19. Hatta, M., Gao, P., Halfmann, P. & Kawaoka, Y. Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses. *Science* **293**, 1773–1775 (2001).
20. Fouchier, R. A. M. *et al.* Avian influenza A virus (H7N7) associated with conjunctivitis and a fatal case of acute respiratory distress syndrome. *Proc. Natl Acad. Sci. USA* **101**, 1356–1361 (2004).
21. Seo, S. H., Hoffmann, E. & Webster, R. G. Lethal H5N1 influenza viruses escape host anti-viral cytokine responses. *Nature Med.* **8**, 950–954 (2002).
22. Peiris, J. S. M. *et al.* Co-circulation of avian H9N2 and contemporary “human” H3N2 influenza viruses in pigs in southeastern China: potential for genetic reassortment? *J. Virol.* **75**, 9679–9686 (2001).
23. Sims, L. D. *et al.* Avian influenza in Hong Kong 1977–2002. *Avian Dis.* **47**, 832–838 (2003).
24. Guan, Y. *et al.* H9N2 influenza viruses possessing H5N1-like internal genomes continue to circulate in poultry in southeastern China. *J. Virol.* **74**, 9372–9380 (2000).
25. Swofford, D. L. *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods)* 4.0 Beta (Sinauer Associates, Sunderland, USA, 2001).
26. Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**, 1244–1245 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We acknowledge K. Stöhr and the World Health Organization for facilitating the study; L. J. Zhang, C. L. Cheung and Y. H. C. Leung for technical assistance; N. Ng and colleagues for provision of computing facilities; and T. M. Ellis, K. Dyrting, W. Wong, P. Li and C. Li of the Department of Agriculture, Fisheries and Conservation of Hong Kong for their support of field work and W. Lim, for virus isolates. We also thank S. Naron for editorial assistance. These studies were supported by a grant from the National Institutes of Health, a grant from The Wellcome Trust, the Ellison Foundation, the Li Ka Shing Foundation, and grants from the Research Grants Council of Hong Kong.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to Y.G. (yguan@hkucc.hku.hk). The sequences reported in this paper have been deposited in GenBank under accession numbers AY651320–AY651758.

Differential activation of the inflammasome by caspase-1 adaptors ASC and Ipaf

Sanjeev Mariathasan¹, Kim Newton¹, Denise M. Monack⁴, Domagoj Vucic¹, Dorothy M. French¹, Wyne P. Lee², Meron Roose-Girma³, Sharon Erickson³ & Vishva M. Dixit¹

¹Molecular Oncology Department, ²Immunology Department, and ³Physiology Department, Genentech, Inc., 1 DNA Way, South San Francisco, California 94080, USA

⁴Department of Microbiology and Immunology, Stanford School of Medicine, Stanford, California 94305, USA

Specific adaptors regulate the activation of initiator caspases; for example, FADD and Apaf-1 engage caspases 8 and 9, respectively¹. The adaptors ASC, Ipaf and RIP2 have each been proposed to regulate caspase-1 (also called interleukin (IL)-1 converting enzyme), which is activated within the 'inflammasome', a complex comprising several adaptors². Here we show the impact of

ASC-, Ipaf- or RIP2-deficiency on inflammasome function. ASC was essential for extracellular ATP-driven activation of caspase-1 in toll-like receptor (TLR)-stimulated macrophages. Accordingly, ASC-deficient macrophages exhibited defective maturation of IL-1 β and IL-18, and ASC-null mice were resistant to lipopolysaccharide-induced endotoxic shock. Furthermore, activation of caspase-1 in response to an intracellular pathogen (*Salmonella typhimurium*) was abrogated severely in ASC-null macrophages. Unexpectedly, Ipaf-deficient macrophages activated caspase-1 in response to TLR plus ATP stimulation but not *S. typhimurium*. Caspase-1 activation was not compromised by loss of RIP2. These data show that whereas ASC is key to caspase-1 activation within the inflammasome, Ipaf provides a special conduit to the inflammasome for signals triggered by intracellular pathogens. Notably, cell death triggered by stimuli that engage caspase-1 was ablated in macrophages lacking either ASC or Ipaf, suggesting a coupling between the inflammatory and cell death pathways.

Excessive levels of the proinflammatory cytokines IL-1 β and IL-18 are associated with septic shock and autoimmune syndromes³. Activated monocytes and macrophages use caspase-1 to cleave pro-IL-1 β and pro-IL-18 to the mature cytokines. Recent studies have suggested that caspase-1 is activated within a multiple adaptor complex termed the inflammasome⁴. ASC (apoptosis-associated speck-like protein containing a CARD)⁵, also known as Pycard⁴ or TMS1 (ref. 6), is one putative component of the inflammasome. A caspase activation and recruitment domain (CARD) within ASC binds to the caspase-1 prodomain, and over-expression studies have reported that ASC both promotes and inhibits activation of caspase-1 (refs 4, 7–9). To establish whether ASC is a positive or negative regulator of caspase-1 in a physiological setting, we generated ASC-deficient mice by gene targeting (Supplementary Fig. S1).

We investigated the role of ASC in caspase-1-dependent IL-1 β release *ex vivo* using thioglycollate-elicited peritoneal macrophages. Synthesis, processing and release of mature IL-1 β by macrophages *ex vivo* seem to require two distinct stimuli. An inflammatory stimulus such as lipopolysaccharide (LPS) primes cells to transcribe and synthesize pro-IL-1 β , and then a second stimulus—such as ATP,

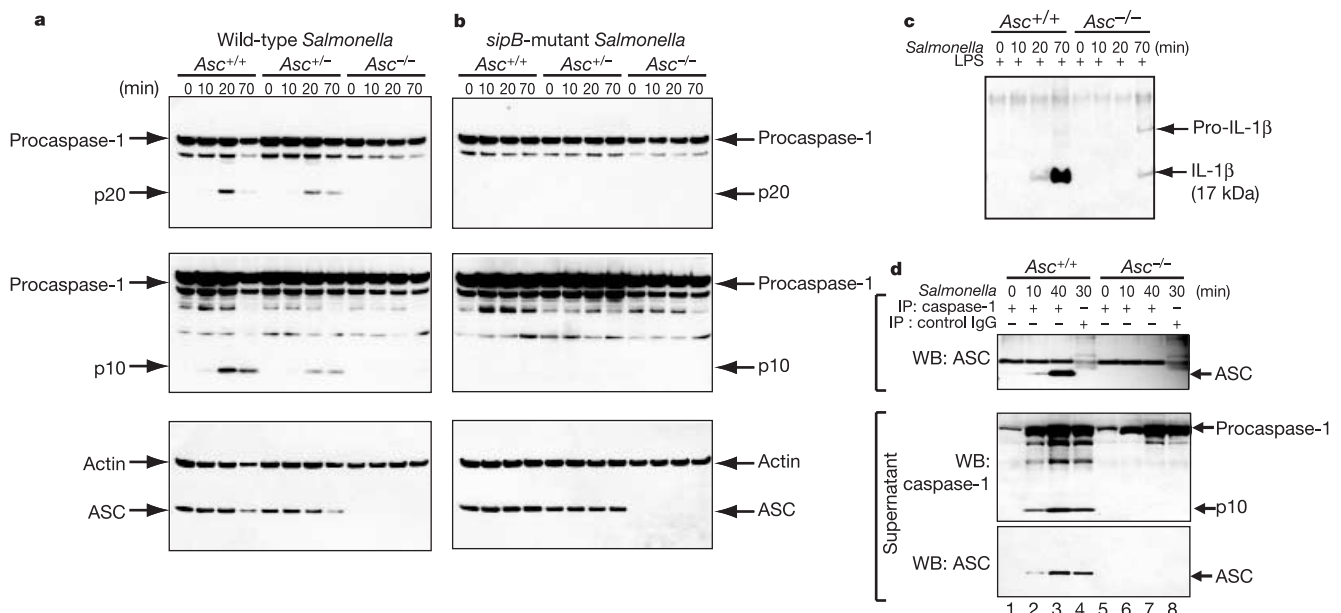


Figure 1 ASC is essential for *S. typhimurium*-induced caspase-1 activation. LPS-stimulated macrophages were infected with either wild-type SL1344 (a, c, d) or a *sipB*-mutant *S. typhimurium* (b) for the times indicated. a, b, Cell lysates were immunoblotted with antibodies against the p10 and p20 subunits of caspase-1, ASC and

actin. c, Pro-IL-1 β and mature IL-1 β in culture supernatants were detected by immunoprecipitation and western blotting. d, Caspase-1 present in culture supernatants was immunoprecipitated with antibodies to the caspase-1 p10 subunit and co-precipitating ASC was detected by western blotting.

nigericin, bacterial toxins (for example, *S. typhimurium* SipB, *Shigella flexneri* IpaB), or cytotoxic T-cell encounter—is necessary for maximal IL-1 β release^{10–13}. In the absence of a second signal, most of the pro-IL-1 β is degraded or remains intracellular and unprocessed. These stimuli probably activate caspase-1 in conjunction with the relevant adaptor proteins. Autocatalytic processing of 45 kDa caspase-1 generates two subunits, p20 and p10 (ref. 14). Thus, LPS-primed macrophages from *Asc*^{+/+}, *Asc*^{+/-} and *Asc*^{-/-} mice were infected with *S. typhimurium* and examined by immunoblotting using antibodies raised against either the p20 or p10 subunit of mouse caspase-1 (Fig. 1a). Both subunits of caspase-1 were evident in *Asc*^{+/+} and *Asc*^{+/-} macrophages within 20 min of infection, but could not be detected in *Asc*^{-/-} macrophages, even at 70 min after infection. Consistent with previous studies¹³, a non-cytotoxic *sipB* mutant of *S. typhimurium* failed to induce caspase-1 cleavage in *Asc*^{+/+}, *Asc*^{+/-} or *Asc*^{-/-} macrophages (Fig. 1b). LPS priming was not essential for optimal caspase-1 activation by wild-type *S. typhimurium*, but was necessary for pro-IL-1 β synthesis (Fig. 2g; see also Supplementary Fig. 2). Next we compared LPS-primed *Asc*^{+/+} and *Asc*^{-/-} macrophages for maturation and release of IL-1 β in response to *S. typhimurium* (Fig. 1c). Western blotting

for IL-1 β after its immunoprecipitation from culture supernatants revealed that *Asc*^{-/-} macrophages secreted much less mature IL-1 β than their wild-type counterparts. These data indicate that ASC is essential for *S. typhimurium*-induced caspase-1 activation.

To see whether ASC associates with caspase-1 in primary macrophages, caspase-1 was immunoprecipitated from LPS-primed *Asc*^{+/+} macrophages infected with *S. typhimurium*. *Asc*^{-/-} macrophages were used as a negative control. We failed to detect any co-precipitating ASC (Supplementary Fig. 3). Notably, infection of wild-type cells with *S. typhimurium* led to the appearance of both caspase-1 and ASC in the culture supernatant. Because caspase-1 subunits have been localized to cell surface membranes¹⁵ and release of IL-1 β from monocytic THP-1 cells has been reported to occur by microvesicle shedding¹⁶, it may be that a complex containing ASC and caspase-1 is shed from the macrophage too rapidly for us to capture it in cell lysates. Indeed, when caspase-1 was immunoprecipitated from culture supernatants, ASC was co-precipitated (Fig. 1d). These results support the notion that ASC associates with caspase-1 to promote its activation, and confirm previous observations of inflammasome components being released from cells secreting mature IL-1 β ⁴.

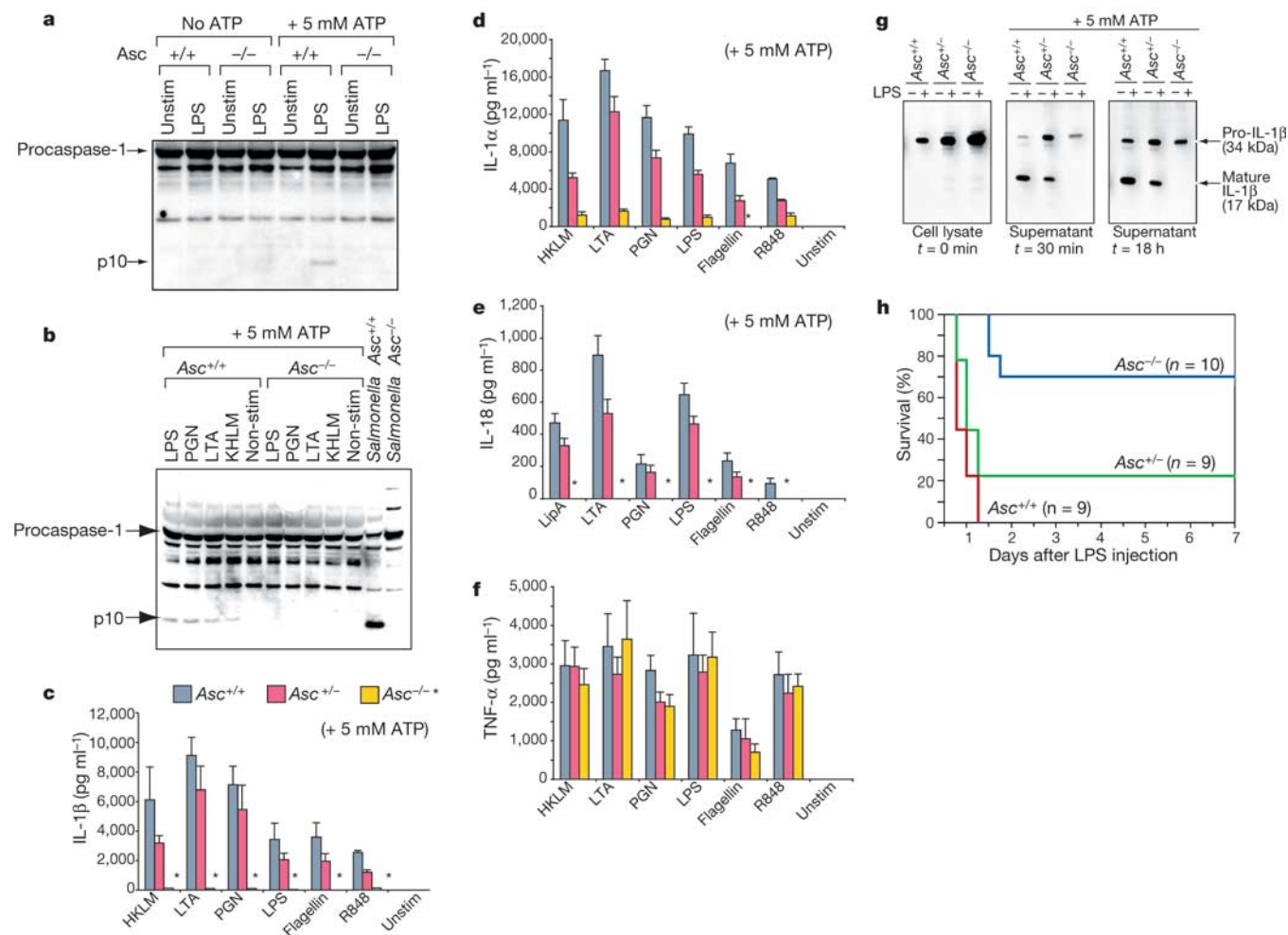


Figure 2 ASC deficiency impairs maturation of IL-1 β and IL-18 in response to TLR agonists and extracellular ATP. **a**, Macrophages were stimulated with or without LPS (500 ng ml⁻¹) for 16 h and then pulsed with ATP for 20 min. Lysates were immunoblotted with antibodies against the p10 subunit of caspase-1. **b–f**, Macrophages were treated with various TLR agonists and were pulsed transiently with ATP for 20 min. **b**, Lysates were immunoblotted with antibodies against the p10 subunit of caspase-1. **c–e**, IL-1 β (**c**), IL-1 α (**d**) and IL-18 (**e**) release was measured by ELISA. **f**, TNF secreted during the initial 16 h incubation with TLR agonists was measured by ELISA. Results are

representative of four independent experiments. Error bars in **c–f** represent the standard deviation of triplicate wells. HKLM, heat-killed *Listeria monocytogenes*; LipA, lipid A; LTA, lipoteichoic acid; PGN, peptidoglycan. **g**, Macrophages were stimulated with or without LPS for 3 h and then pulsed with [³⁵S]-methionine. Cells were lysed (left panel) or treated with ATP and chased for the indicated times. Immunoprecipitation and western blotting was performed using antibodies to IL-1 β . **h**, Mice were injected intraperitoneally with LPS (without additional ATP) and their survival was monitored.

Next we tested $Asc^{+/+}$, $Asc^{+/-}$ and $Asc^{-/-}$ macrophages for caspase-1 activation and IL-1 β release after priming with TLR agonists and treatment with ATP. ATP reportedly triggers P2X₇-purinergic receptors to induce caspase-1-dependent maturation of IL-1 β ^{10,11}. Unlike $Asc^{+/+}$ macrophages, $Asc^{-/-}$ macrophages primed with TLR agonists did not process caspase-1 in response to ATP (Fig. 2a, b). It is noteworthy that ATP alone was insufficient for caspase-1 cleavage in wild-type macrophages (Fig. 2a), whereas TLR priming was dispensable for *S. typhimurium*-induced caspase-1 cleavage (Supplementary Fig. 2). Release of IL-1 β , IL-18 and IL-1 α from ATP-treated, TLR-primed $Asc^{-/-}$ macrophages was also impaired when compared with $Asc^{+/+}$ macrophages (Fig. 2c–e, g; see also Supplementary Fig. 4). Pro-IL-1 α is not a substrate for caspase-1 but a similar defect in IL-1 α secretion was observed for caspase-1 mutant macrophages^{17,18}. $Asc^{+/+}$, $Asc^{+/-}$ and $Asc^{-/-}$ macrophages secreted comparable amounts of the proinflammatory cytokine tumour-necrosis factor (TNF) (Fig. 2f), indicating a specific requirement for ASC in caspase-1-dependent secretion of IL-1 β , IL-1 α and IL-18.

We checked that the defect in IL-1 β secretion by $Asc^{-/-}$ macrophages was not due to inhibition of pro-IL-1 β synthesis by immunoprecipitating pro-IL-1 β and mature IL-1 β from LPS-treated, [³⁵S]-methionine-pulsed macrophages (Fig. 2g).

Comparable amounts of pro-IL-1 β were detected in LPS-treated $Asc^{+/+}$, $Asc^{+/-}$ and $Asc^{-/-}$ macrophages, demonstrating that ASC is dispensable for LPS-induced pro-IL-1 β synthesis. In addition, whereas pro-IL-1 β was secreted into the supernatant by $Asc^{+/+}$, $Asc^{+/-}$ and $Asc^{-/-}$ macrophages subsequent to ATP treatment, only $Asc^{+/+}$ and $Asc^{+/-}$ cells secreted mature IL-1 β (Fig. 2g; see also Supplementary Fig. 5). This result is consistent with a specific role for ASC in caspase-1-dependent cleavage of IL-1 β .

High doses of LPS lead to endotoxic shock and death in mice, presumably due to massive, systemic release of proinflammatory cytokines. Caspase-1 is a crucial mediator of this inflammatory response¹⁷. To determine whether ASC is required for caspase-1 activation *in vivo*, we injected $Asc^{+/+}$, $Asc^{+/-}$ and $Asc^{-/-}$ mice with a lethal dose of 50 mg kg⁻¹ of LPS (Fig. 2h; see also Supplementary Fig. 6). By 12 h, 50% of $Asc^{+/+}$ mice had died, and 100% mortality was observed by 36 h. Survival of $Asc^{-/-}$ mice was enhanced significantly; only 30% of the mice had died after 48 h. Survival of $Asc^{+/-}$ mice was also enhanced relative to wild-type mice (Fig. 2h). Surviving mice exhibited mild lethargy, coat ruffling, febrile shaking and eye watering but had recovered by day 7. Our results demonstrate that ASC, like caspase-1, is an important mediator of LPS-induced endotoxic shock.

Coexpression of ASC with certain other PYRIN-domain-

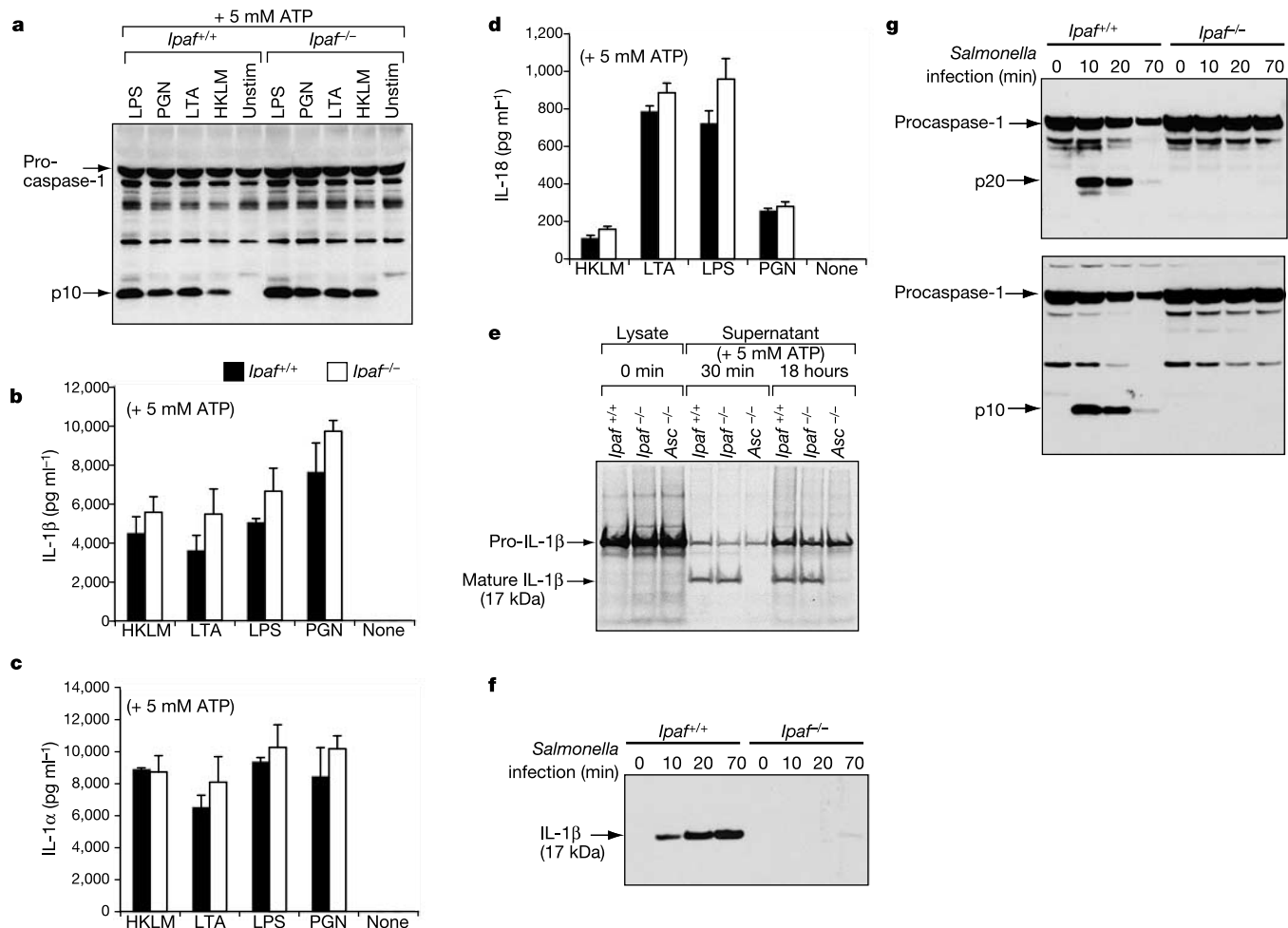


Figure 3 *Ipaf* is essential for caspase-1 activation in response to *S. typhimurium*. **a–d**, Macrophages were stimulated with TLR agonists and pulsed with ATP. **a**, Cell lysates were immunoblotted with antibodies against the p10 subunit of caspase-1. **b–d**, IL-1 β (**b**), IL-1 α (**c**) and IL-18 (**d**) were released into the media and measured by ELISA. Results are representative of four independent experiments. Error bars represent the standard deviation of triplicate wells. **e**, LPS-stimulated macrophages were pulsed

with [³⁵S]-methionine. Cells were lysed (left panel) or treated with ATP and chased for indicated time points. Immunoprecipitations and western blotting were performed with antibodies to IL-1 β . **f**, **g**, LPS-stimulated macrophages were infected with *S. typhimurium* for the times indicated. **f**, Culture supernatants were analysed for IL-1 β by western blotting. **g**, Cell lysates were immunoblotted with antibodies against the p10 or p20 subunits of caspase-1.

containing proteins has been reported to activate NF- κ B-dependent gene expression^{8,19}, whereas small interfering RNA (siRNA)-mediated knockdown of ASC enhances TNF- and LPS-mediated NF- κ B signalling¹⁹. We investigated whether ASC influences LPS-

and TNF-induced NF- κ B signalling by western blot analysis of *Asc*^{+/+} and *Asc*^{-/-} macrophages using antibodies to I κ B α . I κ B proteins sequester NF- κ B family members in the cytosol but get phosphorylated and degraded in response to TNF or LPS. Phosphorylation and then degradation of I κ B α was observed in both *Asc*^{+/+} and *Asc*^{-/-} macrophages on TNF or LPS treatment (Supplementary Fig. 7). Phosphorylation of the mitogen-activated kinases ERK1 and ERK2 was also normal in *Asc*^{-/-} macrophages. Therefore we find no evidence of a critical role for ASC in either ERK or NF- κ B signalling in response to TNF or LPS.

Several other adaptors besides ASC have been proposed to interact with caspase-1 and modulate its activity. These proteins include RIP2 (receptor-interacting protein 2) and Ipaf (ICE-protease-activating factor)^{20–23}. Although overexpression studies suggest that RIP2 induces caspase-1 activation, no defects were reported for RIP2-deficient mice with respect to TLR-induced IL-1 β processing and release²⁴. However, a role for RIP2 in *S. typhimurium*-induced caspase-1 activation was not examined. We measured IL-1 β and IL-18 released from *Rip2*^{+/+} and *Rip2*^{-/-} macrophages in response to either ATP and TLR agonists or *S. typhimurium*. RIP2 deficiency had no impact on the secretion of these cytokines (Supplementary Fig. 8), indicating that RIP2 is dispensable for normal caspase-1-dependent maturation of IL-1 β and IL-18.

Ipaf²³ (also known as CARD12 (ref. 22) and CLAN (ref. 25)) has an amino-terminal CARD, and overexpression studies have suggested that it associates with either caspase-1 (refs 22, 23, 25) or ASC²². To determine the role of Ipaf in caspase-1 activation, we generated Ipaf-deficient mice by gene targeting (Supplementary Fig. 9). Thioglycollate-elicited macrophages from *Ipaf*^{+/+} and *Ipaf*^{-/-} mice exhibited equivalent caspase-1 processing in response to TLR stimulation and extracellular ATP (Fig. 3a; see also Supplementary Fig. 4c) and secreted comparable amounts of IL-1 β , IL-1 α and IL-18 (Fig. 3b–e). *Ipaf*^{-/-} macrophages also produced normal amounts of IL-6 and TNF in response to TLR signalling (data not shown); however, negligible amounts of IL-1 β were produced by *Ipaf*^{-/-} macrophages infected with *S. typhimurium* when compared with *Ipaf*^{+/+} macrophages (Fig. 3f). Western blotting for the p10 and p20 subunits of caspase-1 indicated that caspase-1 processing was absent in the infected *Ipaf*^{-/-} cells (Fig. 3g) despite ASC being present at normal levels (Supplementary Fig. 10). These results indicate that Ipaf and ASC are both essential for caspase-1 activation in response to the intracellular pathogen *S. typhimurium*. How Ipaf might engage the inflammasome remains to be determined.

Cells lacking caspase-1 respond normally to most apoptotic stimuli¹⁷ with one notable exception being macrophages infected by *S. flexneri* or *S. typhimurium*; in the early stages of infection, a significant proportion of macrophages are killed in a caspase-1-dependent manner^{12,13,26}. To test whether this cell death is dependent on ASC and Ipaf, the viability of wild-type, *Ipaf*^{-/-} and *Asc*^{-/-} macrophages infected with *S. typhimurium* was monitored by lactate dehydrogenase (LDH) release. At 20 min after infection, 60% of wild-type cells were dead compared with only 7% of *Asc*^{-/-} cells and less than 1% of *Ipaf*^{-/-} cells (Fig. 4a). Maximum caspase-1 cleavage was observed in wild-type macrophages at 10 min after infection (Fig. 4b) but, in agreement with our earlier results (Figs 1a and 3g), only the zymogen form of caspase-1 was detected in *Ipaf*^{-/-} and *Asc*^{-/-} macrophages (Fig. 4b). Conversely, *sipB*-mutant *S. typhimurium* did not kill or promote caspase-1 processing in wild-type, *Asc*^{-/-} or *Ipaf*^{-/-} macrophages (Fig. 4c, d). By 2 h after infection, cell death in cultures of wild-type and *Asc*^{-/-} macrophages was comparable (Fig. 4a). These findings are consistent with late-stage cell death triggered by *S. typhimurium* being independent of caspase-1 activation²⁷. Interestingly, even at 2 h after infection, *Ipaf*^{-/-} macrophages were markedly resistant to *S. typhimurium*-induced cell death; approximately 90% of cells remained viable by this assay (Fig. 4a). The increased viability of *Ipaf*^{-/-} macrophages

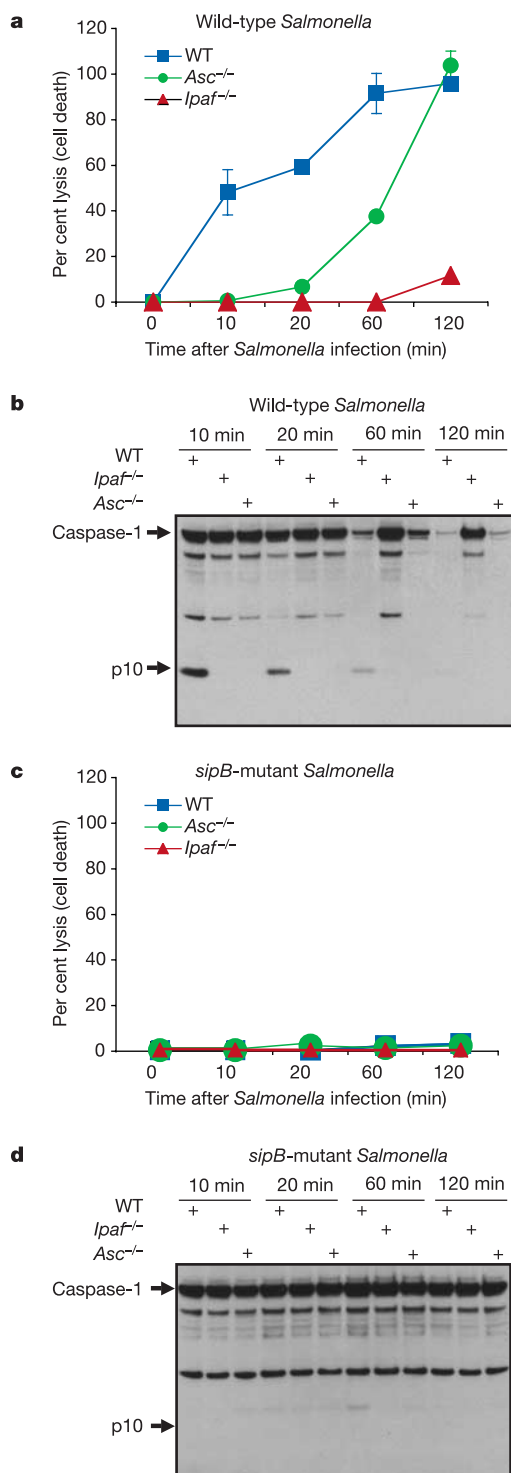


Figure 4 ASC- and Ipaf-deficient macrophages are resistant to cell death induced by *S. typhimurium* infection. Macrophages from wild-type (WT), *Asc*^{-/-} and *Ipaf*^{-/-} mice were infected with either wild-type SL1344 *S. typhimurium* (**a**, **b**) or *sipB*-mutant *S. typhimurium* (**c**, **d**). **a**, **c**, Cell death was measured by LDH release into the culture supernatant and is shown as a percentage of LDH release by detergent. Data points represent the mean $\pm$ standard deviation of triplicate wells. Results are representative of two independent experiments. **b**, **d**, Cell lysates were western blotted with antibodies against the p10 subunit of caspase-1.

relative to $Asc^{-/-}$ macrophages at 2 h after infection suggests that Ipaf might also engage signalling pathways distinct from those leading to activation of the inflammasome.

Our results provide genetic evidence that ASC is essential for caspase-1 activation within the inflammasome and that Ipaf is essential for caspase-1 activation in response to at least one intracellular pathogen. Future studies will address the relationship between Ipaf and ASC. The role of disease-associated proteins that contain a PYRIN domain and appear to interact with ASC, such as pyrin and cryopyrin/Nalp3, also needs to be determined^{28,29}. Also implicated in inflammasome function is Nalp1 (ref. 4), a protein that contains both a PYRIN domain and a CARD. Analyses of mice lacking these adaptors should reveal whether they are an essential component of the inflammasome, like ASC, or serve to transduce signals to the inflammasome from distinct upstream stimuli, like Ipaf. □

Methods

Asc, Ipaf and Rip2 mutant mice

The generation of RIP2-deficient mice has been described³⁰. *Asc* (Supplementary Fig. 1a) and *Ipaf* (Supplementary Fig. 9a) targeting vectors were electroporated into 129 R1 and C57BL/6 C2 embryonic stem (ES) cells, respectively. Heterozygous ES cell clones were identified by Southern blotting and were microinjected into C57BL/6N (*Asc*) or BALB/c blastocysts (*Ipaf*). Chimaeric offspring were backcrossed to C57BL/6N mice and germline transmission was confirmed by polymerase chain reaction and Southern blot analysis of tail DNA.

TLR agonists and antibodies

Heat-killed *Listeria monocytogenes* (10^8 bacteria per ml), lipoteichoic acid (100 ng ml⁻¹) from *Bacillus subtilis*, peptidoglycan (10 µg ml⁻¹) from *Staphylococcus aureus*, LPS (100 ng ml⁻¹) from *Escherichia coli*, and R-848 (10 ng ml⁻¹) were purchased from InvivoGen. Flagellin (100 µg ml⁻¹) was purchased from Calbiochem and lipid A from *E. coli* was purchased from Sigma. Recombinant mTNF-α was from PeproTech. The following antibodies were used for western blotting and immunoprecipitations: rat anti-caspase-1 (clone 4B4 raised against amino acids 104–294 of mouse caspase-1; Genentech); rabbit anti-caspase-1 (sc-514; Santa Cruz Biotechnology); rabbit or rat anti-ASC (raised against full-length mouse ASC; Genentech); rabbit or hamster anti-*Ipaf* (raised against amino acids 610–826 of mouse *Ipaf*; Genentech); mouse anti-actin (ICN Biomedicals); hamster anti-mouse IL-1β (clone B122; Becton Dickinson); goat anti-mouse IL-1β (clone AF-401-NA; R&D Systems). Phospho-IκBα (Ser 32), IκBα, phospho-ERK and ERK antibodies were from Cell Signalling Technology.

Infection with *S. typhimurium*

Mice were injected intraperitoneally with 4% thioglycollate solution and macrophages were collected by peritoneal lavage 5 days later. A total of 2×10^6 cells were plated in a six-well dish and non-adherent cells were removed after 2 h. Adherent macrophages were cultured in the high glucose version of Dulbecco's modified Eagle's medium supplemented with 10% heat-inactivated fetal calf serum, 50 µg ml⁻¹ penicillin, 50 µg ml⁻¹ streptomycin and 500 ng ml⁻¹ LPS overnight. Cells were washed with fresh culture media that lacked antibiotics and infected with either wild-type *S. typhimurium* SL1344 or *sipB*-mutant *S. typhimurium* at a multiplicity of infection of 50.

Immunoprecipitations and western blotting

Macrophages infected with *S. typhimurium* were lysed in RIPA buffer (10 mM phosphate buffer pH 7.4, 150 mM NaCl, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS) supplemented with Complete protease inhibitor cocktail (Roche) and 2 mM dithiothreitol. Lysates were resolved in 4–12% Tris-glycine gradient gels (Invitrogen) and transferred to nitrocellulose (Invitrogen) by electro-blotting. For immunoprecipitations, lysates were pre-cleared with protein A/G agarose beads (Pierce). Culture supernatants from macrophages infected with *S. typhimurium* were mixed with an equal volume of 2X RIPA buffer before immunoprecipitation with goat anti-mouse IL-1β (clone AF-401-NA; R&D Systems) and immunoblotted with hamster anti-mouse IL-1β (clone B122; Becton Dickinson). For caspase-1 activation assays with TLR and ATP, cells and media supernatants were lysed together.

Cytokine ELISA measurements

Thioglycollate-elicited peritoneal macrophages were stimulated with TLR agonists for 16 h and then culture supernatants were assayed for TNF and IL-6. The TLR-stimulated macrophages were cultured for a further 20 min in fresh medium containing 5 mM ATP (Sigma), washed, and then cultured in fresh medium for an additional 3 h. Culture supernatants then were assayed for IL-1β, IL-α and IL-18. Enzyme-linked immunosorbent assay (ELISA) kits were from R&D Systems.

Pulse-chase analyses

Peritoneal macrophages were treated with or without LPS (1 µg ml⁻¹) for 3 h and pulse-labelled with [³⁵S]-methionine (200 µCi ml⁻¹; ICN Biomedicals) for 45 min. Labelled cells were washed with PBS and treated with ATP (5 mM) for 30 min, then the medium was

collected and fresh medium added. Culture supernatants were collected after a 3 or 18 h chase. Immunoprecipitations with purified hamster anti-mouse IL-1β antibodies (clone B122; Becton Dickinson) and autoradiography were performed as previously described¹⁰.

LPS-induced septic shock

Mice (8–10 weeks old) were injected intraperitoneally with 50 mg kg⁻¹ of LPS from *E. coli* (serotype 0111:B4; Sigma). The mice were monitored for signs of endotoxaemia and lethality twice daily for 5 days and periodically thereafter. No extracellular ATP was injected for this survival study. For serum cytokine measurements, another cohort of mice received 50 mg kg⁻¹ of LPS and the animals were bled retro-orbitally 3.5 h after LPS injection. These mice received an additional 50 µl intraperitoneal injection of 100 mM ATP 15 min before bleeding to induce maximal IL-1β release.

Macrophage death assays

Macrophages were seeded in 96-well plates at 5×10^4 cells per well and incubated overnight with 500 ng ml⁻¹ LPS. Before infection with *S. typhimurium*, the medium was replaced with serum-free RPMI 1640 medium. Cells were infected with either wild-type *S. typhimurium* SL1344 or *sipB*-mutant *S. typhimurium* grown in high salt and low O₂ with a multiplicity of infection of 50. Cultures were supplemented with gentamycin (90 µg ml⁻¹) after 30 min to kill extracellular bacteria. Cell death at the time points indicated was quantified colorimetrically with the CytoTox96 LDH-release kit (Promega) as previously described¹³.

Received 24 March; accepted 17 May 2004; doi:10.1038/nature02664.

Published online 9 June 2004.

- Ashkenazi, A. & Dixit, V. M. Death receptors: signaling and modulation. *Science* **281**, 1305–1308 (1998).
- Tschopp, J., Martinon, F. & Burns, K. NALPs: a novel protein family involved in inflammation. *Nature Rev. Mol. Cell Biol.* **4**, 95–104 (2003).
- Dinarello, C. A. Interleukin-1β, interleukin-18, and the interleukin-1β converting enzyme. *Ann. NY Acad. Sci.* **856**, 1–11 (1998).
- Martinon, F., Burns, K. & Tschopp, J. The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of proIL-1β. *Mol. Cell* **10**, 417–426 (2002).
- Masumoto, J. et al. ASC, a novel 22-kDa protein, aggregates during apoptosis of human promyelocytic leukemia HL-60 cells. *J. Biol. Chem.* **274**, 33835–33838 (1999).
- Conway, K. E. et al. TMS1, a novel proapoptotic caspase recruitment domain protein, is a target of methylation-induced gene silencing in human breast cancers. *Cancer Res.* **60**, 6236–6242 (2000).
- Srinivasula, S. M. et al. The PYRIN-CARD protein ASC is an activating adaptor for caspase-1. *J. Biol. Chem.* **19**, 19–22 (2002).
- Wang, L. et al. PYPAF7, a novel PYRIN-containing Apaf1-like protein that regulates activation of NF-κB and caspase-1-dependent cytokine processing. *J. Biol. Chem.* **277**, 29874–29880 (2002).
- Stehlik, C. et al. Apoptosis-associated speck-like protein containing a caspase recruitment domain is a regulator of procaspase-1 activation. *J. Immunol.* **171**, 6154–6163 (2003).
- Hogquist, K. A., Nett, M. A., Unanue, E. R. & Chaplin, D. D. Interleukin 1 is processed and released during apoptosis. *Proc. Natl Acad. Sci. USA* **88**, 8485–8489 (1991).
- Perregaux, D. & Gabel, C. A. Interleukin-1β maturation and release in response to ATP and nigericin. Evidence that potassium depletion mediated by these agents is a necessary and common feature of their activity. *J. Biol. Chem.* **269**, 15195–15203 (1994).
- Chen, Y., Smith, M. R., Thirumalai, K. & Zychlinsky, A. A bacterial invasin induces macrophage apoptosis by binding directly to ICE. *EMBO J.* **15**, 3853–3860 (1996).
- Hersh, D. et al. The *Salmonella* invasin *SipB* induces macrophage apoptosis by binding to caspase-1. *Proc. Natl Acad. Sci. USA* **96**, 2396–2401 (1999).
- Thornberry, N. A. et al. A novel heterodimeric cysteine protease is required for interleukin-1β processing in monocytes. *Nature* **356**, 768–774 (1992).
- Singer, I. I. et al. The interleukin-1β-converting enzyme (ICE) is localized on the external cell surface membranes and in the cytoplasmic ground substance of human monocytes by immuno-electron microscopy. *J. Exp. Med.* **182**, 1447–1459 (1995).
- MacKenzie, A. et al. Rapid secretion of interleukin-1β by microvesicle shedding. *Immunity* **15**, 825–835 (2001).
- Li, B. et al. Mice deficient in IL-1β-converting enzyme are defective in production of mature IL-1β and resistant to endotoxic shock. *Cell* **80**, 401–411 (1995).
- Kuida, K. et al. Altered cytokine export and apoptosis in mice deficient in interleukin-1β converting enzyme. *Science* **267**, 2000–2003 (1995).
- Stehlik, C. et al. The PAAD/PYRIN-family protein ASC is a dual regulator of a conserved step in nuclear factor κB activation pathways. *J. Exp. Med.* **196**, 1605–1615 (2002).
- Humke, E. W., Shriver, S. K., Starovastnik, M. A., Fairbrother, W. J. & Dixit, V. M. ICEBERG: a novel inhibitor of interleukin-1β generation. *Cell* **103**, 99–111 (2000).
- Thome, M. et al. Identification of CARDIAC, a RIP-like kinase that associates with caspase-1. *Curr. Biol.* **8**, 885–888 (1998).
- Geddes, B. J. et al. Human CARD12 is a novel CED4/Apaf-1 family member that induces apoptosis. *Biochem. Biophys. Res. Commun.* **284**, 77–82 (2001).
- Poyet, J. L. et al. Identification of Ipaf, a human caspase-1-activating protein related to Apaf-1. *J. Biol. Chem.* **276**, 28309–28313 (2001).
- Kobayashi, K. et al. RICK/Rip2/CARDIAC mediates signalling for receptors of the innate and adaptive immune systems. *Nature* **416**, 194–199 (2002).
- Damiano, J. S., Stehlik, C., Pio, F., Godzik, A. & Reed, J. C. CLAN, a novel human CED-4-like gene. *Genomics* **75**, 77–83 (2001).
- Monack, D. M. et al. *Salmonella* exploits caspase-1 to colonize Peyer's patches in a murine typhoid model. *J. Exp. Med.* **192**, 249–258 (2000).
- Jesenberger, V., Procyk, K. J., Yuan, J., Reipert, S. & Baccarini, M. *Salmonella*-induced caspase-2 activation in macrophages: a novel mechanism in pathogen-mediated apoptosis. *J. Exp. Med.* **192**, 1035–1046 (2000).
- Gumucio, D. L. et al. Fire and ICE: the role of pyrin domain-containing proteins in inflammation and apoptosis. *Clin. Exp. Rheumatol.* **S0**, S45–S53 (2002).

29. Inohara, N. & Nunez, G. NODs: intracellular proteins involved in inflammation and apoptosis. *Nature Rev. Immunol.* **3**, 371–382 (2003).
 30. Ruefli-Brasse, A. A., Lee, W. P., Hurst, S. & Dixit, V. M. Rip2 participates in Bcl10 signaling and T-cell receptor-mediated NF- κ B activation. *J. Biol. Chem.* **279**, 1570–1574 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank D. Dornan and other members of the Dixit laboratory for discussions, E. Humke for help with illustrations, and K. O'Rourke, D. Wadley, Z. Gu, C. Olsson, M. Bauer, L. Tom, J. Kloss, M. Fuentes, M. Osborn, C. Tan, J. Hongo, T. Wong and A. Chuntharapai for technical assistance.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to V.M.D. (dixit@genec.com).

Regulation of Toll/IL-1-receptor-mediated gene expression by the inducible nuclear protein I κ B ζ

Masahiro Yamamoto¹, Soh Yamazaki³, Satoshi Uematsu¹, Shintaro Sato^{1,2}, Hiroaki Hemmi¹, Katsuaki Hoshino⁶, Tsuneyasu Kaisho⁶, Hirotaka Kuwata¹, Osamu Takeuchi^{1,2}, Koichiro Takeshige³, Tatsuya Saitoh⁷, Shoji Yamaoka⁷, Naoki Yamamoto⁷, Shunsuke Yamamoto⁸, Tatsushi Muta^{3,4}, Kiyoshi Takeda⁵ & Shizuo Akira^{1,2}

¹Department of Host Defense, Research Institute for Microbial Diseases, Osaka University, ²ERATO, Japan Science and Technology Agency, 3-1 Yamada-oka, Suita Osaka 565-0871, Japan

³Department of Molecular and Cellular Biochemistry, Graduate School of Medical Sciences, ⁴PRESTO, Japan Science and Technology Agency, ⁵Department of Embryonic and Genetic Engineering, Medical Institute of Bioregulation, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka 812-8582, Japan

⁶RIKEN Research Center for Allergy and Immunology, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

⁷Department of Molecular Virology, Graduate School of Medicine, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8519, Japan

⁸Department of Food and Nutrition, Beppu University, Kita-ishigaki, Beppu, Oita 874-0851, Japan

Toll-like receptors (TLRs) recognize microbial components and trigger the inflammatory and immune responses against pathogens. I κ B ζ (also known as MAIL and INAP) is an ankyrin-repeat-containing nuclear protein that is highly homologous to the I κ B family member Bcl-3 (refs 1–6). Transcription of I κ B ζ is rapidly induced by stimulation with TLR ligands and interleukin-1 (IL-1). Here we show that I κ B ζ is indispensable for the expression of a subset of genes activated in TLR/IL-1R signalling pathways. I κ B ζ -deficient cells show severe impairment of IL-6 production in response to a variety of TLR ligands as well as IL-1, but not in response to tumour-necrosis factor- α . Endogenous I κ B ζ specifically associates with the p50 subunit of NF- κ B, and is recruited to the NF- κ B binding site of the IL-6 promoter on stimulation. Moreover, NF- κ B1/p50-deficient mice show responses to TLR/IL-1R ligands similar to those of I κ B ζ -deficient mice. Endotoxin-induced expression of other genes such as *Il12b* and *Csf2* is also abrogated in I κ B ζ -deficient macrophages. Given that the lipopolysaccharide-induced transcription of I κ B ζ occurs earlier than transcription of these genes, some TLR/IL-1R-mediated responses may be regulated in a gene expression process of at least two steps that requires inducible I κ B ζ .

I κ B ζ is thought to be induced in response to IL-1 or lipopolysaccharide (LPS) (TLR4 ligand)^{4–6}. In addition to IL-1 and LPS,

I κ B ζ messenger RNA was strongly upregulated on stimulation with peptidoglycan (PGN) (TLR2 ligand), bacterial lipoprotein (BLP) (TLR1/TLR2), flagellin (TLR5), MALP-2 (TLR6/TLR2), R-848 (TLR7) and CpG DNA (TLR9), but not with tumour-necrosis factor- α (TNF- α) (Fig. 1a). In contrast, other I κ B family members such as I κ B α and Bcl-3 were induced in response to TNF- α as well as the TLR ligands and IL-1. Thus, I κ B ζ is induced in the TLR/IL-1R signalling pathway but not the TNF signalling pathway. Furthermore, IL-1- or LPS-induced expression of I κ B ζ was completely abolished in *MyD88*^{−/−} embryonic fibroblasts (MEFs; Fig. 1b), showing that I κ B ζ is inducible in the MyD88-dependent part of the TLR/IL-1R signalling pathway^{7–15}.

To elucidate the physiological role of I κ B ζ in the TLR/IL-1R response, we generated *I κ B ζ* ^{−/−} mice by targeted gene disruption (see Supplementary Discussion 1 and Supplementary Fig. 1a–d). *I κ B ζ* ^{−/−} splenocytes showed defective proliferation in response to LPS but not to anti-CD40, IL-4 and anti-IgM (Supplementary Fig. 1e, f), suggesting that the TLR4 response is impaired in *I κ B ζ* ^{−/−} cells. Moreover, although *I κ B ζ* ^{−/−} mice grew normally after birth, some of them started to develop atopic dermatitis-like skin lesions with acanthosis and lichenoid changes at the age of 4–5 weeks (Supplementary Fig. 2a, b). All *I κ B ζ* ^{−/−} mice developed the disease by the age of 10 weeks. Histological analysis of 5-week-old *I κ B ζ* ^{−/−} mice showed pathological changes in the conjunctiva, including a heavy lymphocyte infiltration into the submucosa and loss of goblet cells in the conjunctival epithelium (Supplementary Fig. 2c–f).

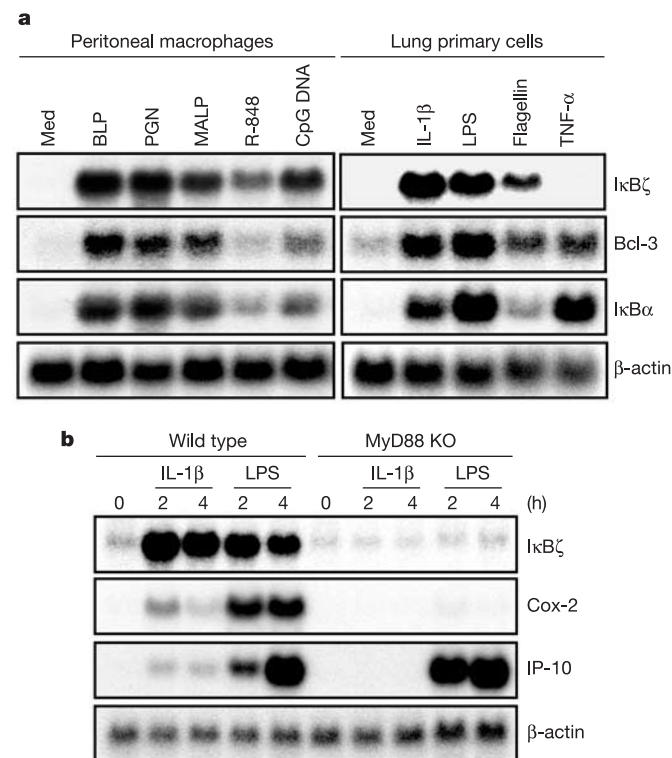


Figure 1 Specific induction of I κ B ζ on stimulation by TLR/IL-1R ligands. **a**, The indicated cells were stimulated with 100 ng ml^{−1} BLP, 10 μ g ml^{−1} PGN, 30 ng ml^{−1} MALP-2, 100 nM R-848, 3 μ M CpG DNA, 10 ng ml^{−1} IL-1 β , 10 μ g ml^{−1} LPS, 100 ng ml^{−1} flagellin and 10 ng ml^{−1} TNF- α for 2 h. Total RNA (10 μ g) was extracted and subjected to northern blot analysis for expression of I κ B ζ , Bcl-3, I κ B α and β -actin. Med, medium. **b**, *MyD88*^{+/+} (wild type) and *MyD88*^{−/−} (KO) MEFs were stimulated with 10 ng ml^{−1} IL-1 β and 10 μ g ml^{−1} LPS for the indicated periods. Total RNA (10 μ g) was extracted and subjected to northern blot analysis for expression of I κ B ζ , Cox-2, IP-10 and β -actin.

We analysed the LPS-induced production of inflammatory mediators in macrophages. *IκBζ*^{+/+} macrophages produced TNF-α, IL-6 and nitric oxide (NO) in response to LPS (Fig. 2a). Although LPS-induced production of TNF-α and NO was normal, production of IL-6 was severely impaired in *IκBζ*^{-/-} macrophages. In addition, production of IL-6 via stimulation with various TLR ligands was also profoundly inhibited in *IκBζ*^{-/-} cells (Fig. 2b–d). Moreover, *IκBζ*^{-/-} cells exhibited defective IL-1-induced IL-6 production; however, IL-1-induced activation of NF-κB and mitogen-activated protein kinases was not impaired in these cells, indicating that there is no defect in the intracellular signalling pathways (Supplementary Fig. 3a, b). On the other hand, TNF-α-induced IL-6 production was not impaired in *IκBζ*^{-/-} cells (Fig. 2e). The impaired production of IL-6 in response to LPS correlated well with the reduced induction of IL-6 mRNA in *IκBζ*^{-/-} cells (Fig. 2f).

When full-length or a deletion mutant form of *IκBζ* was transfected into MEFs, full-length *IκBζ*, but not the deletion mutant,

rescued the defective production of IL-6 on stimulation with IL-1 in *IκBζ*^{-/-} cells (Fig. 2g), suggesting that expression of *IκBζ* is required for TLR/IL-1-mediated production of IL-6. As the genes for *IκBζ* and IL-6 are inducible in response to TLR ligands and IL-1 (refs 4–6, 16, 17), we compared the time course of mRNA expression in macrophages. On stimulation with LPS, induction of *IκBζ* expression was observed at 30 min and reached maximal levels after 120 min. On the other hand, induction of IL-6 mRNA occurred at later time points compared with *IκBζ* or TNF-α (Fig. 2h). Taken together, these results indicate that the TLR/IL-1R-mediated expression of the IL-6 gene (*Il6*) requires the preceding induction of *IκBζ*. Given that *IκBζ* is also an inducible protein in TLR/IL-1R-mediated signalling pathways, the TLR/IL-1R-mediated production of pro-inflammatory IL-6 may be controlled in at least a two-step fashion.

Our data and those of a previous study⁵ indicate the positive role of *IκBζ* in the TLR/IL-1R-mediated expression of IL-6. To test

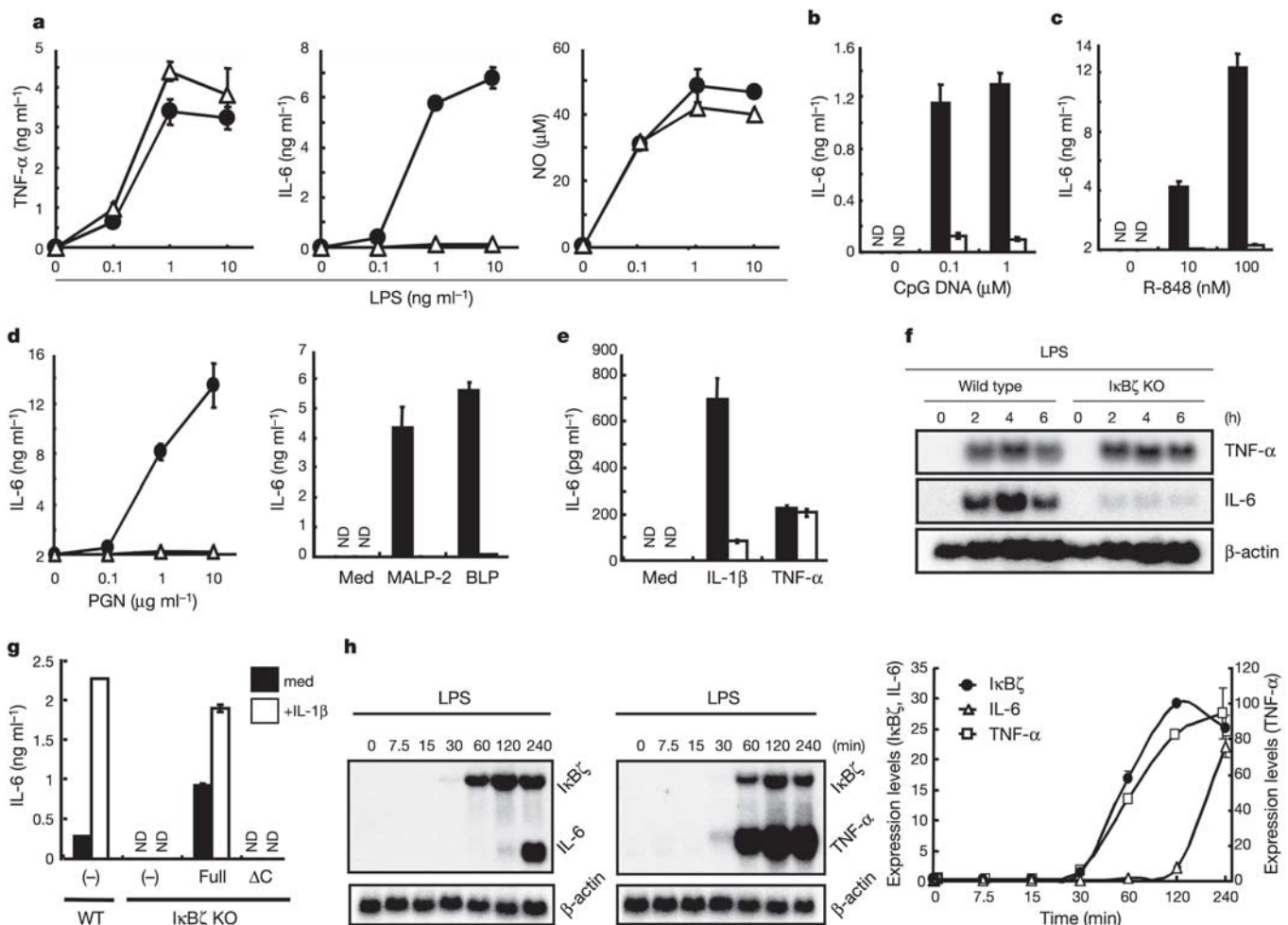


Figure 2 Immune responses in *IκBζ*^{-/-} cells and kinetics of *IκBζ* induction. **a–d**, *IκBζ*^{+/+} (filled symbols/columns) and *IκBζ*^{-/-} (open symbols/columns) peritoneal macrophages were cultured with 10 ng ml⁻¹ LPS, 100 ng ml⁻¹ BLP, 30 ng ml⁻¹ MALP-2, and the indicated concentrations of PGN, R-848 and CpG DNA in the presence of 30 ng ml⁻¹ IFN-γ for 24 h. Values are means ± s.d. of triplicate experiments. ND, not detected. **e**, *IκBζ*^{+/+} (filled columns) and *IκBζ*^{-/-} (open columns) MEFs were stimulated with 10 ng ml⁻¹ IL-1β and 10 ng ml⁻¹ TNF-α. Values are means ± s.d. of triplicate experiments. **f**, Peritoneal macrophages were stimulated with 10 ng ml⁻¹ LPS for the indicated periods. Total RNA (5 μg) was extracted and subjected to northern blot analysis

for expression of IL-6, TNF-α and β-actin. **g**, Rescue of IL-1 responsiveness in *IκBζ*^{-/-} MEFs by retroviral transfection with full-length (Full), but not deletion mutant (ΔC), *IκBζ*. Indicated values are means ± s.d. of triplicate experiments. **h**, Double RNA products indicative of *IκBζ*, IL-6 and TNF-α mRNA transcripts after LPS stimulation of wild-type peritoneal macrophages resolved by electrophoresis. Two independent experiments with independently derived wild-type cells were quantified by PhosphorImager (left and middle panels), and mRNA abundance (right panel) is shown for the indicated genes in arbitrary units (left axis, *IκBζ* and IL-6; right axis, TNF-α) relative to β-actin. Indicated values are means ± s.d. of duplicate experiments.

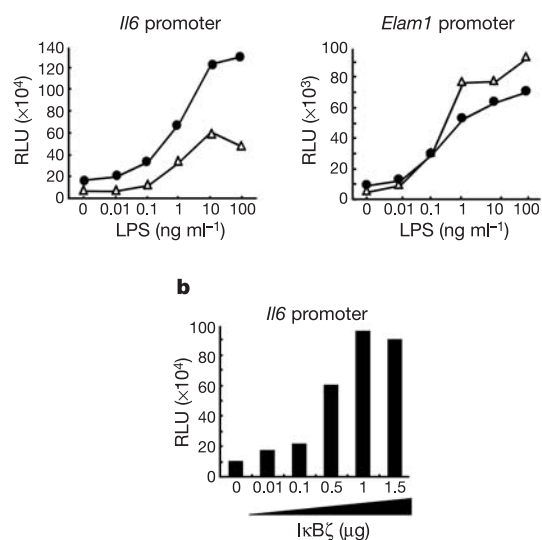
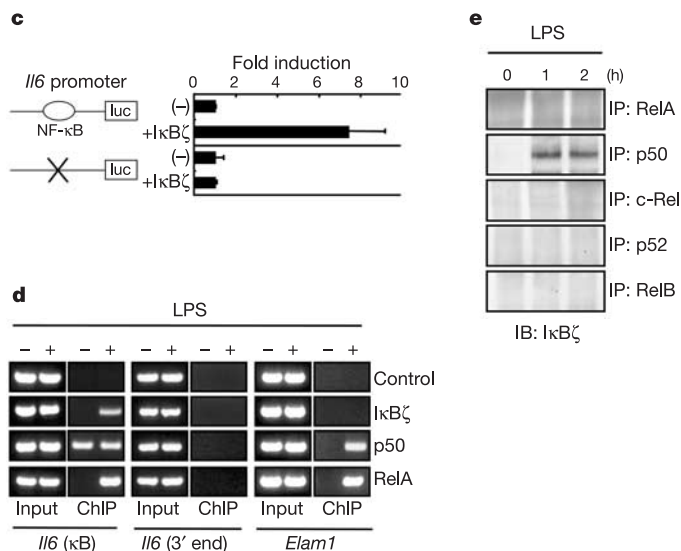


Figure 3 *In vitro* analysis of IκBζ on the *I/6* promoter. **a**, RAW 264.7 cells were transiently transfected with luciferase reporter constructs of either the murine *I/6* promoter or the *Elam1* promoter together with either control (open symbols) or the IκBζ expression plasmid (filled symbols). Luciferase activities are expressed as fold-increase values over the background shown by lysates prepared from untransfected cells. Data are representative of three separate experiments. Cells were stimulated with the indicated concentrations of LPS. RLU, relative luciferase units. **b**, Untreated RAW 264.7 cells were transiently transfected with the IκBζ expression vector together with constant amounts of the *I/6* reporter plasmid. Data are representative of three separate experiments. **c**, P19 cells were transiently transfected with either wild-type or mutant *I/6* promoter reporter constructs together with either control or the IκBζ expression plasmid. Luciferase



activities were normalized in each case by dividing the fold-increase values of IκBζ-expressed cells over the background values of lysates with that of mock-expressed cells. Values are means $\pm$ s.d. of triplicate experiments. **d**, Chromatin from untreated (–) or LPS-treated (+) (1 μg ml⁻¹ for 3 h) RAW 264.7 cells was used for ChIP assays with the indicated antibodies. Precipitated DNA for the *I/6* κB site (left), the 3' region of the *I/6* gene (centre), or the *Elam1* promoter (right) was assayed by PCR (ChIP). Data are representative of two independent experiments. **e**, Unstimulated or LPS-stimulated (10 ng ml⁻¹) peritoneal macrophages were immunoprecipitated with the indicated antibodies. The immunoprecipitated (IP) lysates were subsequently immunoblotted (IB) with anti-IκBζ.

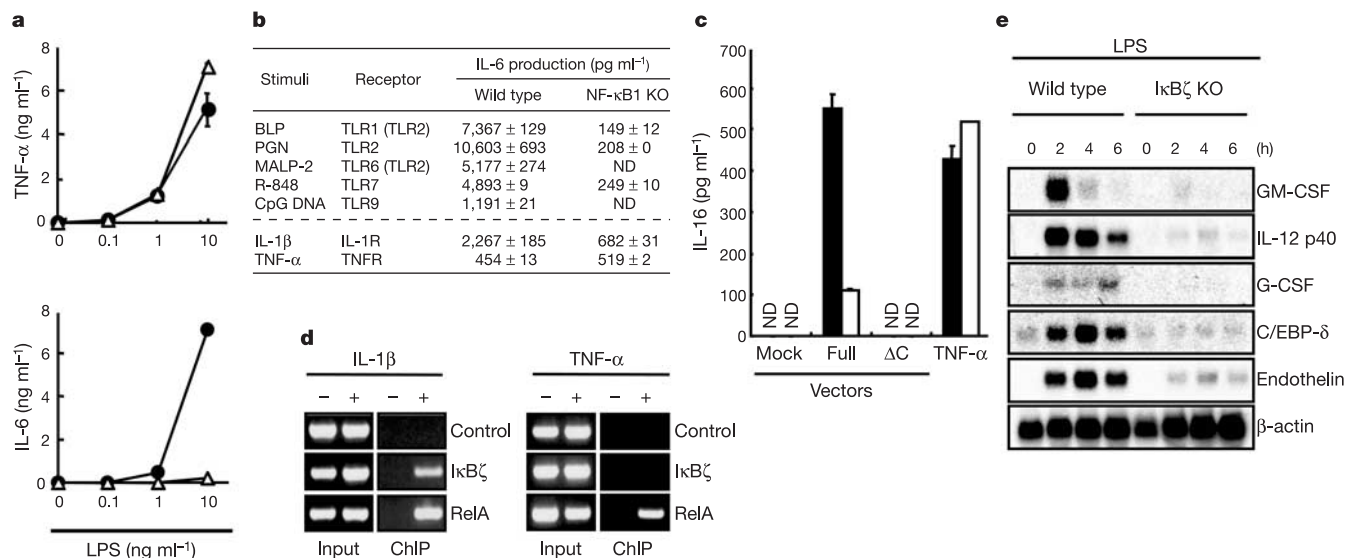


Figure 4 The TLR/IL-1R responses in NF-κB1/p50-deficient cells and microarray analysis of *IκBζ*^{-/-} cells. **a**, *NF-κB1*^{+/+} (filled symbols) and *NF-κB1*^{-/-} (open symbols) peritoneal macrophages were cultured with 10 ng ml⁻¹ LPS in the presence of 30 ng ml⁻¹ IFN-γ for 24 h. Values are means $\pm$ s.d. of triplicate experiments. **b**, *NF-κB1*^{+/+} and *NF-κB1*^{-/-} peritoneal macrophages and MEFs were cultured with 100 ng ml⁻¹ BLP, 10 μg ml⁻¹ PGN, 30 ng ml⁻¹ MALP-2, 100 nM R-848, 3 μM CpG DNA, 10 ng ml⁻¹ IL-1β or 10 ng ml⁻¹ TNF-α in the presence of 30 ng ml⁻¹ IFN-γ for 24 h. IL-1β- and TNF-α-induced IL-6 production were analysed by MEFs. Values are means $\pm$ s.d. of triplicate experiments. **c**, *NF-κB1*^{+/+} (filled columns) and *NF-κB1*^{-/-} (open columns) MEFs were retrovirally transfected with either the full-length (Full) or the

deletion mutant (Δ C) of IκBζ. Furthermore, the same lines of untransfected cells were stimulated with 10 ng ml⁻¹ TNF-α. Values are means $\pm$ s.d. of duplicate experiments. **d**, Chromatin from untreated (–), IL-1β-treated (left panel +); 10 ng ml⁻¹ for 3 h) or TNF-α-treated (right panel +); 10 ng ml⁻¹ for 3 h) wild-type MEFs were used for ChIP assays with the indicated antibodies. Precipitated DNA for the input (left) or the *I/6* κB site (right) was assayed by PCR. **e**, *IκBζ*^{+/+} and *IκBζ*^{-/-} peritoneal macrophages were stimulated with 10 ng ml⁻¹ LPS for the indicated periods. Total RNA (5 μg) was extracted and subjected to northern blot analysis for expression of the indicated probes. GM-CSF, granulocyte-macrophage colony-stimulating factor; G-CSF, granulocyte colony-stimulating factor.

whether I κ B ζ promotes the ligand-induced activation of the *Il6* promoter, we introduced reporter plasmids containing the promoter of either *Il6* or *Elam1* (also known as *Sele*) into RAW 264.7 cells together with control or I κ B ζ expression vectors. LPS stimulation activated both the *Il6* and *Elam1* promoters in a dose-dependent manner. Under conditions of I κ B ζ overexpression, we found that the promoter activities of *Il6*, but not *Elam1*, were further enhanced (Fig. 3a). Ectopic expression of I κ B ζ alone also upregulated the activity of the *Il6* promoter in unstimulated cells (Fig. 3b). As an NF- κ B binding site in the *Il6* promoter has been shown to have an important function in its activation, we transfected P19 cells with either a wild-type reporter or a mutant reporter in which the NF- κ B binding site was disrupted¹⁸. I κ B ζ overexpression activated the wild-type reporter, but not the mutant NF- κ B reporter (Fig. 3c). To probe directly the specific involvement of I κ B ζ in the κ B site of the *Il6* promoter, we performed a chromatin immunoprecipitation assay (ChIP) to investigate the proteins bound to the region. In unstimulated cells, the NF- κ B p50 subunit was readily detected in the *Il6* κ B site as described previously¹⁹. On stimulation with LPS, RelA as well as p50 bound to the κ B sites of the *Il6* and the *Elam1* promoter, but not the 3' end of the *Il6* gene. In contrast, we found that I κ B ζ only bound to the *Il6* κ B site, but not the other sites tested, in LPS-stimulated cells (Fig. 3d), demonstrating a specificity of I κ B ζ for the κ B site in the *Il6* promoter. Finally, we addressed association of I κ B ζ with NF- κ B family members. Immunoprecipitation analysis

showed that I κ B ζ proteins interacted with p50, but not with RelA, RelB, c-Rel or the p52 subunit (Fig. 3e). These findings indicate that the positive effects by I κ B ζ may be exerted through association with the p50 subunit.

The aforementioned results prompted us to study NF- κ B/p50-deficient cells. As previously reported, whereas LPS-induced TNF- α production in NF- κ B1^{-/-} macrophages is normal, NF- κ B1^{-/-} macrophages show defective LPS-induced production of IL-6 (refs 20, 21; see also Fig. 4a). Additionally, production of IL-6 in response to stimulation by IL-1 and other TLR ligands was severely impaired in NF- κ B1^{-/-} cells (Fig. 4b). Thus, the TLR/IL-1R-mediated responses in NF- κ B1^{-/-} mice are similar to those in I κ B ζ ^{-/-} mice. We tested further whether I κ B ζ overexpression induced IL-6 production in NF- κ B1^{-/-} cells. Compared with NF- κ B1^{+/+} cells, I κ B ζ -mediated production of IL-6 was markedly reduced in NF- κ B1^{-/-} MEFs, indicating that NF- κ B1 is critical for the effect of I κ B ζ (Fig. 4c). Both IL-1 and TNF- α activate similar sets of signalling molecules such as NF- κ B and mitogen-activated protein kinases, and culminate in IL-6 expression. However, signalling mediated by IL-1/TLR ligands, but not TNF- α , specifically recruited I κ B ζ to the *Il6* promoter (Fig. 4d), presumably accounting for the responsiveness difference between TLR/IL-1R and TNFR. Next, we searched for other LPS-inducible genes regulated by I κ B ζ using microarray analysis. Dozens of LPS-inducible genes were significantly affected by the I κ B ζ deficiency (Supplementary Fig. 4a). Several of them were subsequently tested by northern blot analysis for confirmation of accuracy. Among them, LPS-induced expression of granulocyte-macrophage colony-stimulating factor (*Csf2*), IL-12 p40 (*Il12b*), granulocyte colony-stimulating factor (*Csf3*), C/EBP- δ (*Cebpd*) and endothelin 1 (*Edn1*) was compromised in I κ B ζ ^{-/-} macrophages (Fig. 4e). Indeed, time course northern blot analysis showed that the kinetics of LPS induction of these genes were similar to that for IL-6 rather than TNF- α (Fig. 2h, Supplementary Fig. 4b, and data not shown), further supporting the model of a two-step regulation of these genes in TLR/IL-1R signalling pathways. Regarding the relationship between the κ B sequences of the genes tested and I κ B ζ requirement, the promoters of I κ B ζ -regulated genes contained distinct κ B sequences. Therefore, it may be difficult to determine whether an arbitrary LPS-inducible gene is I κ B ζ -dependent through simple sequence comparison of the κ B sites (ref. 22; see also Supplementary discussion 2 and Supplementary Fig. 4c-e).

Finally, we examined *in vivo* cytokine production after LPS injection. Although LPS-induced IL-12 p40 production was impaired, IL-6 production was almost normal in I κ B ζ ^{-/-} mice. Surprisingly, I κ B ζ ^{-/-} mice exhibited more prolonged TNF- α production than I κ B ζ ^{+/+} mice (Fig. 5a). As TNF- α is a major IL-6 producer under conditions of endotoxin-induced shock, we next attempted to negate biological activities of TNF- α by prior treatment with anti-TNF- α neutralizing antibodies (anti-TNF NABs) (ref. 23). In anti-TNF NAB-treated I κ B ζ ^{-/-} mice the serum concentration of LPS-induced IL-6 was significantly reduced compared with anti-TNF NAB-treated I κ B ζ ^{+/+} mice (Fig. 5b), demonstrating that the prolonged TNF- α production might compensate for impaired IL-6 production in I κ B ζ ^{-/-} mice. The prolonged TNF- α production might be secondary to the loss of I κ B ζ -regulated factors that negatively modulate TNF- α production. Alternatively, I κ B ζ might act directly as a negative regulator for TNF- α production in certain cells. Although the molecular mechanism of the prolonged TNF- α production remains unknown, such prolonged TNF- α production might lead to development of the skin lesion in aged I κ B ζ ^{-/-} mice as demonstrated in TNF- α -mediated inflammatory diseases occurring in other mouse models^{24,25}.

We provide genetic evidence that I κ B ζ is essential for TLR/IL-1R-mediated IL-6 production. As I κ B ζ itself is an inducible protein, TLR/IL-1R-mediated IL-6 expression may be regulated in a two-

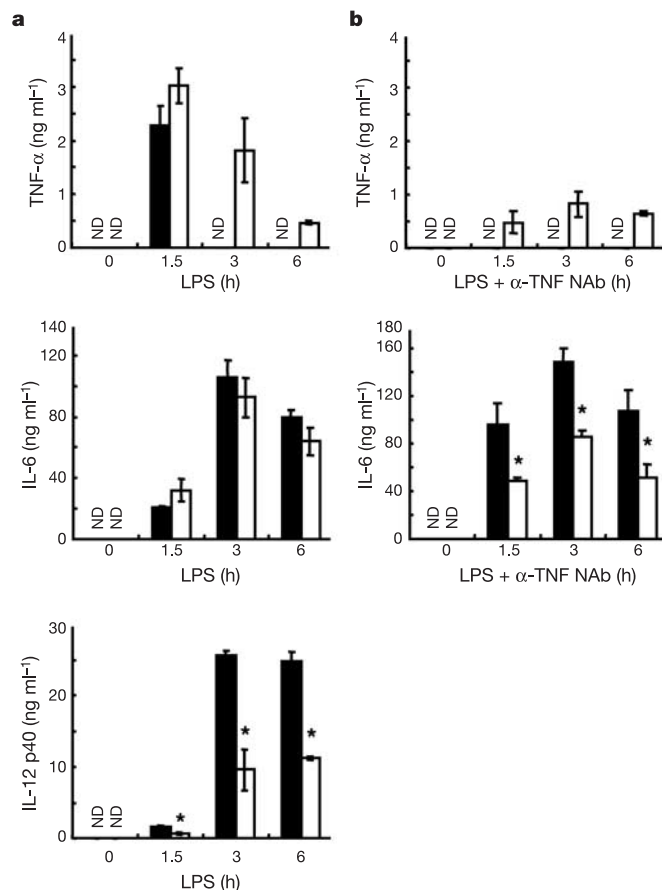


Figure 5 *In vivo* cytokine production in I κ B ζ ^{-/-} mice. **a, b**, Age-matched I κ B ζ ^{-/-} (open columns; $n = 3$ in **a** and 2 in **b**) and I κ B ζ ^{+/+} (filled columns; $n = 3$ in **a** and **b**) mice were intraperitoneally injected with 1.5 mg LPS only (**a**), or with 10 μ g anti-TNF NAB 1 h before the LPS injection (**b**). Sera were collected at the indicated times. Values are means $\pm$ s.d. of sera samples at the indicated times. Asterisk, statistical significance ($P < 0.05$) in a two-tailed Student's *t*-test comparing I κ B ζ ^{+/+} and I κ B ζ ^{-/-} mice.

step mechanism. Moreover, microarray analysis showed that $\text{IkB}\zeta$ might control LPS-inducible genes other than *Il6*. Further analysis that clarifies the molecular basis of $\text{IkB}\zeta$ -dependent gene expression will provide new insight into the TLR/IL-1R-mediated MyD88-dependent immune responses. □

Methods

Generation of $\text{IkB}\zeta^{-/-}$ mice

Genomic DNA containing the $\text{IkB}\zeta$ gene was isolated, as described previously²⁶. We constructed the targeting vector by replacing a 2.0-kilobase (kb) fragment encoding the central portion of $\text{IkB}\zeta$ with a *neo^R* cassette that was transfected into embryonic stem cells (E14.1). G418 and gancyclovir doubly resistant colonies were screened by polymerase chain reaction (PCR) and Southern blotting. We micro-injected two independent homologous recombinants into C57BL/6 blastocysts and intercrossed heterozygous F₁ progenies to obtain $\text{IkB}\zeta^{-/-}$ mice. Mice from these independent clones displayed identical phenotypes. $\text{IkB}\zeta^{-/-}$ mice and their wild-type littermates at the age of 6–12 weeks were used for the current studies. All animal experiments were conducted with the approval of the Animal Research Committee of the Research Institute for Microbial Diseases (Osaka University, Osaka, Japan).

Reagents and mice

LPS, PGN, MALP-2, flagellin, R-848 and CpG oligodeoxynucleotides were prepared as described previously²⁶. Polyclonal antibody against $\text{IkB}\zeta$ was obtained by immunizing rabbit with the C-terminal portion of the murine $\text{IkB}\zeta$ protein. We used antibodies against phosphorylated ERK and JNK (Cell signalling), antibodies against ERK, JNK, p38, RelA, p50, c-Rel, RelB, p52 (Santa Cruz), antibodies against RelA in ChIP assays (Biomol) and antibodies against murine TNF- α for neutralization (R&D). Recombinant TNF- α and IL-1 β were obtained from Genzyme. NF- κ B1/p50-deficient mice were as described previously²⁰.

Measurement of pro-inflammatory cytokine and NO

Thioglycollate-elicited peritoneal macrophages or MEFs were cultured as described previously²⁶. Concentrations of TNF- α (Genzyme), IL-6 (R&D) and IL-12 p40 (Genzyme) in the culture supernatant were measured by an enzyme-linked immunosorbent assay according to the manufacturer's instructions. Concentrations of NO were measured by the Griess method according to the manufacturers' instructions (DOJINDO). To measure *in vivo* cytokine concentrations, sera were taken from $\text{IkB}\zeta^{+/+}$ or $\text{IkB}\zeta^{-/-}$ mice. Anti-TNF NABs were reconstituted in PBS to 20 $\mu\text{g ml}^{-1}$, and 10 μg (500 μl) of the reagent was intraperitoneally injected 1 h before LPS injection.

Electrophoretic mobility shift assay

This assay was performed as described previously²⁶.

Plasmids and retroviral transfection

The reporter plasmids consisted of the 5' flanking region (−1240/+40) of the murine *Il6* gene, and were used in Fig. 3a, b. The reporter plasmids used in Fig. 3c were as described previously^{18,27}. The full-length or the deletion mutant $\text{IkB}\zeta$, which lacks the C-terminal portion (Δ C; ref. 4), was cloned into pMRX retroviral vector, and the transfection was performed as described previously²⁸.

Luciferase reporter assay

Reporter plasmids were transiently co-transfected into RAW 264.7 and P19 cells with either the control or $\text{IkB}\zeta$ expression vectors using SUPERFECT transfection reagent (Qiagen). Luciferase activities of total cell lysates were measured using the Dual-luciferase reporter assay system (Promega) as described previously²⁹.

Chromatin immunoprecipitation assay

The ChIP assay was performed essentially with a described protocol (Upstate Biotechnology). 2×10^6 RAW 264.7 cells or 5×10^5 MEFs were stimulated with LPS (1 $\mu\text{g ml}^{-1}$ for 3 h), IL-1 β (10 ng ml^{-1} for 3 h) or TNF- α (10 ng ml^{-1} for 3 h), respectively. Precipitated DNAs were analysed by quantitative PCR (35–40 cycles) using primers 5'-CGATGCTAAACGACGTACATTGTGCA-3' and 5'-CTCCAGAGCAGAATGAGCTACAGACAT-3' for the κ B site in the *Il6* promoter, 5'-GCAGATGGACTTAGCTCGTCTCATTCA-3' and 5'-CCACTCCTTCTGTGACTCCAGCTTATC-3' for a 3' gene segment in the *Il6* promoter and 5'-GATGCAGTTGAGAATTTCTCTTAGCC-3' and 5'-TGGAATAGTTGTTCTGGCGTTGGATCC-3' for the κ B site in the *Elam1* promoter.

Western blot analysis and immunoprecipitation

Western blot was performed as described previously²⁶.

Gene chip analysis

Microarray analysis (Affimetrix) using $\text{IkB}\zeta^{+/+}$ and $\text{IkB}\zeta^{-/-}$ peritoneal macrophages was performed as described previously³⁰. The colour image for gene expression was generated by GeneSpring6.0 (Silicon Genetics) software.

Histological analysis

Tissues were fixed in 10% phosphate-buffered formalin, and paraffin-embedded tissue sections were stained with haematoxylin and eosin or PAS staining using standard techniques.

Received 1 April; accepted 7 June 2004; doi:10.1038/nature02738.

1. Takeda, K., Kaisho, T. & Akira, S. Toll-like receptors. *Annu. Rev. Immunol.* **21**, 335–376 (2003).
2. Janeway, C. A. Jr & Medzhitov, R. Innate immune recognition. *Annu. Rev. Immunol.* **20**, 197–216 (2002).
3. O'Neill, L. A. Therapeutic targeting of Toll-like receptors for inflammatory and infectious diseases. *Curr. Opin. Pharmacol.* **3**, 396–403 (2003).
4. Yamazaki, S., Muta, T. & Takeshige, K. A novel IkB protein, $\text{IkB}\zeta$, induced by proinflammatory stimuli, negatively regulates nuclear factor- κ B in the nuclei. *J. Biol. Chem.* **276**, 27657–27662 (2001).
5. Kitamura, H., Kanehira, K., Okita, K., Morimatsu, M. & Saito, M. MAIL, a novel nuclear IkB protein that potentiates LPS-induced IL-6 production. *FEBS Lett.* **485**, 53–56 (2000).
6. Haruta, H., Kato, A. & Todokoro, K. Isolation of a novel interleukin-1-inducible nuclear protein bearing ankyrin-repeat motifs. *J. Biol. Chem.* **276**, 12485–12488 (2001).
7. Muzio, M., Ni, J., Feng, P. & Dixit, V. M. IRAK (Pelle) family member IRAK-2 and MyD88 as proximal mediators of IL-1 signaling. *Science* **278**, 1612–1615 (1997).
8. Wesche, H., Henzel, W. J., Shillinglaw, W., Li, S. & Cao, Z. MyD88: an adapter that recruits IRAK to the IL-1 receptor complex. *Immunity* **7**, 837–847 (1997).
9. Medzhitov, R. *et al.* MyD88 is an adaptor protein in the hToll/IL-1 receptor family signaling pathways. *Mol. Cell* **2**, 253–258 (1998).
10. Alexopoulos, L., Holt, A. C., Medzhitov, R. & Flavell, R. A. Recognition of double-stranded RNA and activation of NF- κ B by Toll-like receptor 3. *Nature* **413**, 732–738 (2001).
11. Kawai, T. *et al.* Lipopolysaccharide stimulates the MyD88-independent pathway and results in activation of IFN-regulatory factor 3 and the expression of a subset of lipopolysaccharide-inducible genes. *J. Immunol.* **167**, 5887–5894 (2001).
12. Doyle, S. *et al.* IRF3 mediates a TLR3/TLR4-specific antiviral gene program. *Immunity* **17**, 251–263 (2002).
13. Ghosh, S. & Karin, M. Missing pieces in the NF- κ B puzzle. *Cell* **109** (suppl.), S81–S96 (2002).
14. Zhang, G. & Ghosh, S. Molecular mechanisms of NF- κ B activation induced by bacterial lipopolysaccharide through Toll-like receptors. *J. Endotoxin Res.* **6**, 453–457 (2000).
15. Janssens, S. & Beyaert, R. Functional diversity and regulation of different interleukin-1 receptor-associated kinase (IRAK) family members. *Mol. Cell* **11**, 293–302 (2003).
16. Eto, A., Muta, T., Yamazaki, S. & Takeshige, K. Essential roles for NF- κ B and a Toll/IL-1 receptor domain-specific signal(s) in the induction of $\text{IkB}\zeta$. *Biochem. Biophys. Res. Commun.* **301**, 495–501 (2003).
17. Akira, S., Taka, T. & Kishimoto, T. Interleukin-6 in biology and medicine. *Adv. Immunol.* **54**, 1–78 (1993).
18. Matsusaka, T. *et al.* Transcription factors NF-IL6 and NF- κ B synergistically activate transcription of the inflammatory cytokines, interleukin 6 and interleukin 8. *Proc. Natl Acad. Sci. USA* **90**, 10193–10197 (1993).
19. Zhong, H., May, M. J., Jimi, E. & Ghosh, S. The phosphorylation status of nuclear NF- κ B determines its association with CBP/p300 or HDAC-1. *Mol. Cell* **9**, 625–636 (2002).
20. Sha, W. C., Liou, H. C., Tuomanen, E. I. & Baltimore, D. Targeted disruption of the p50 subunit of NF- κ B leads to multifocal defects in immune responses. *Cell* **80**, 321–330 (1995).
21. Sanjabi, S., Hoffmann, A., Liou, H. C., Baltimore, D. & Smale, S. T. Selective requirement for c-Rel during IL-12 p40 gene induction in macrophages. *Proc. Natl Acad. Sci. USA* **97**, 12705–12710 (2000).
22. Hoffmann, A., Leung, T. H. & Baltimore, D. Genetic analysis of NF- κ B/Rel transcription factors defines functional specificities. *EMBO J.* **22**, 5530–5539 (2003).
23. Hill, M. R. & McCallum, R. E. Identification of tumor necrosis factor as a transcriptional regulator of the phosphoenolpyruvate carboxykinase gene following endotoxin treatment of mice. *Infect. Immun.* **60**, 4040–4050 (1992).
24. Pasparakis, M. *et al.* TNF-mediated inflammatory skin disease in mice with epidermis-specific deletion of IKK2. *Nature* **417**, 861–866 (2002).
25. Douni, E. *et al.* Transgenic and knockout analyses of the role of TNF in immune regulation and disease pathogenesis. *J. Inflamm.* **47**, 27–38 (1995).
26. Yamamoto, M. *et al.* Role of adaptor TRIF in the MyD88-independent Toll-Like Receptor signaling pathway. *Science* **301**, 640–643 (2003).
27. Kinoshita, S., Akira, S. & Kishimoto, T. A member of the C/EBP family, NF-IL6 beta, forms a heterodimer and transcriptionally synergizes with NF-IL6. *Proc. Natl Acad. Sci. USA* **89**, 1473–1476 (1992).
28. Saitoh, T. *et al.* TWEAK induces NF- κ B2 p100 processing and long lasting NF- κ B activation. *J. Biol. Chem.* **278**, 36005–36012 (2003).
29. Toshchakov, V. *et al.* TLR4, but not TLR2, mediates IFN- β -induced STAT1 α/β -dependent gene expression in macrophages. *Nature Immunol.* **3**, 392–398 (2002).
30. Kuwata, H. *et al.* IL-10-inducible Bcl-3 negatively regulates LPS-induced TNF- α production in macrophages. *Blood* **102**, 4123–4129 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank T. Kitamura, A. Aderem, D. Golenbock and H. Tomizawa for providing Plat-E packaging cell lines, flagellin, the ELAM1 reporter plasmid, and R-848, respectively. We also thank T. Kawai and K. Ishii for discussions; M. Hashimoto for secretarial assistance; and N. Okita and N. Iwami for technical assistance. This work was supported by grants from Special Coordination Funds, the Ministry of Education, Culture, Sports, Science and Technology, Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists, The Uehara Memorial Foundation, The Naito Foundation, and The Junior Research Associate from RIKEN.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.A. (sakira@biken.osaka-u.ac.jp).

Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors

Anton Meinhart & Patrick Cramer

Department of Chemistry and Biochemistry, Gene Center, University of Munich, 81377 Munich, Germany

During transcription, RNA polymerase (Pol) II synthesizes eukaryotic messenger RNA. Transcription is coupled to RNA processing by the carboxy-terminal domain (CTD) of Pol II, which consists of up to 52 repeats of the sequence Tyr 1-Ser 2-Pro 3-Thr 4-Ser 5-Pro 6-Ser 7 (refs 1, 2). After phosphorylation, the CTD binds tightly to a conserved CTD-interacting domain (CID) present in the proteins Pcf11 and Nrd1, which are essential and evolutionarily conserved factors for polyadenylation-dependent and -independent 3'-RNA processing, respectively. Here we describe the structure of a Ser 2-phosphorylated CTD peptide bound to the CID domain of Pcf11. The CTD motif Ser 2-Pro 3-Thr 4-Ser 5 forms a β -turn that binds to a conserved groove in the CID domain. The Ser 2 phosphate group does not make direct contact with the CID domain, but may be recognized indirectly because it stabilizes the β -turn with an additional hydrogen bond. Iteration of the peptide structure results in a compact β -spiral model of the CTD. The model suggests that, during the mRNA transcription-processing cycle, compact spiral regions in

the CTD are unravelled and regenerated in a phosphorylation-dependent manner.

Efficient capping, splicing and polyadenylation of mRNAs *in vivo* requires the CTD of Pol II^{1,2}. The CTD is linked flexibly to the Pol II core close to the presumed RNA exit site³, and binds to mRNA-processing factors *in vitro*^{1,2}. The binding of processing factors is regulated by phosphorylation of CTD residues Ser 2 and Ser 5. During transcription initiation, the CTD becomes phosphorylated at Ser 5 (ref. 4), which triggers binding of the 5'-capping enzyme for the first RNA-processing event. Subsequent transcription elongation is accompanied by CTD phosphorylation at Ser 2 residues⁴, which is required for efficient 3'-RNA processing^{5,6}. Thus, distinct CTD phosphorylation patterns seem to couple different phases of the transcription cycle to the corresponding RNA-processing events, suggesting that a CTD code specifies the position of Pol II in the cycle⁷. During or after transcription termination, the CTD is dephosphorylated, resulting in Pol II recycling.

Several RNA-processing factors recognize the CTD by means of a conserved CTD-interacting domain (CID; Fig. 1a). Factors with CID domains include the serine/arginine-rich-like factors SCAF4 and SCAF8 (refs 8, 9), Nrd1 (which is implicated in polyadenylation-independent RNA 3'-end formation¹⁰) and Pcf11. Pcf11 is a conserved and essential subunit of the yeast cleavage factor IA, which is required for 3'-RNA processing and transcription termination¹¹. The high-resolution X-ray structure of the Pcf11 CID domain was determined (see Methods and Supplementary Information). The structure identified eight α -helices in a right-handed superhelical arrangement (Fig. 1b). The CID fold closely resembles that of VHS domains¹² and is related to armadillo-

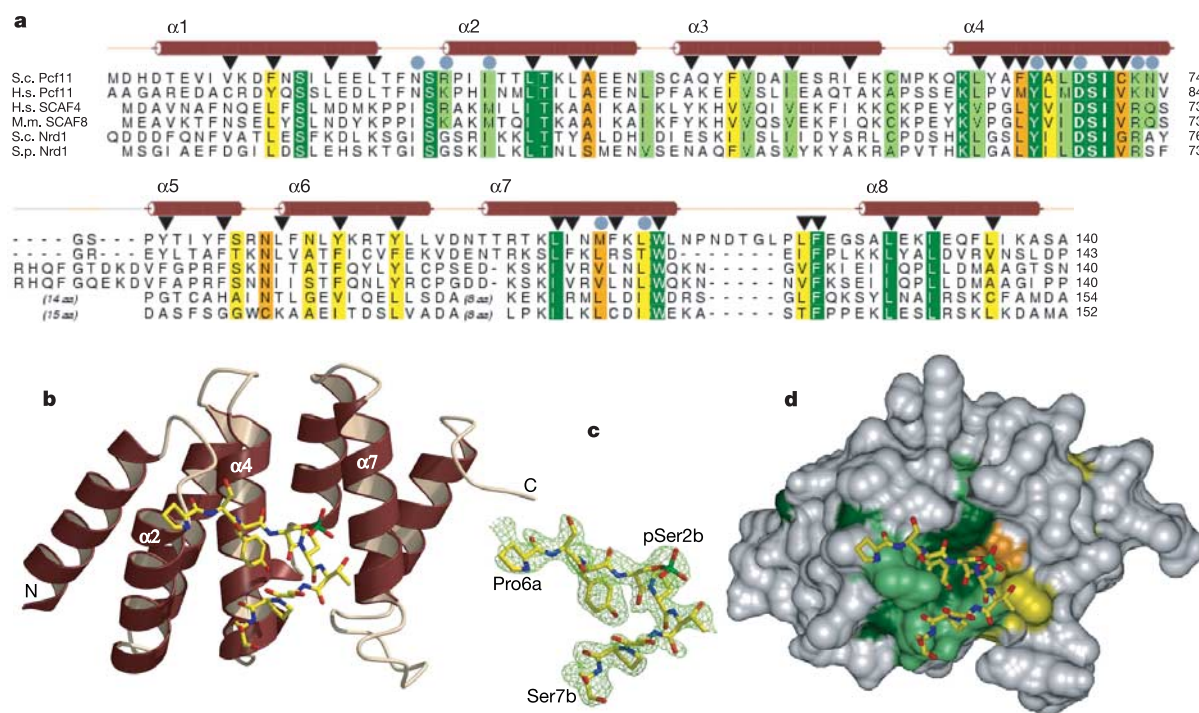


Figure 1 Structure of the CID-CTD complex. **a**, Alignment of amino acid sequences of CID domains in *S. cerevisiae* (S.c.) Pcf11, human (H.s.) Pcf11, human SCAF4, mouse (M.m.) SCAF8, *S. cerevisiae* Nrd1 and *Schizosaccharomyces pombe* (S.p.) Nrd1. α -Helices are indicated as cylinders above the alignment, and residues involved in hydrophobic-core packing are marked with a triangle. Residues involved in CTD binding are marked with a dot. Conserved residues are highlighted with the degree of conservation decreasing from

dark green to yellow. **b**, Structure of the CID-CTD complex. The Pcf11 CID domain is shown as a ribbon model in dark red and the CTD phosphopeptide is shown as an atomic model. **c**, Simulated annealing omit electron density map for the CTD peptide at 2.2 Å resolution, contoured at 3 σ . **d**, CTD peptide binding to a conserved groove in the CID domain. Conserved surface residues are colored according to **a**.

repeat proteins¹³, except for the two amino-terminal helices. Amino acid residues in the hydrophobic core of the domain are highly conserved across CID domains (Fig. 1a), demonstrating that the structure is a valid model for all of these domains.

To determine how the CID domain recognizes the CTD, we soaked into CID crystals a synthetic CTD peptide with the following sequence: Pro 3a-Thr 4a-Ser 5a-Pro 6a-Ser 7a-Tyr 1b-pSer 2b-Pro 3b-Thr 4b-Ser 5b-Pro 6b-Ser 7b, which comprises a central phosphorylated Ser 2 residue (pSer2b). A difference Fourier map produced detailed electron density for all peptide residues, except for the three N-terminal residues (Fig. 1c). The CTD peptide bound to a groove formed by α -helices 2, 4 and 7 (Fig. 1b). As the groove is lined by conserved residues (Fig. 1d), other CID domains may use the same CTD-binding site. The crystal lattice does not seem to impose any constraints on peptide binding, as the conserved groove is fully solvent-exposed and is not involved in crystal packing (see Methods).

The CTD peptide shows a β -turn at its central motif (pSer 2b-Pro 3b-Thr 4b-Ser 5b), whereas the flanking residues are in an extended conformation (Fig. 2). The CTD backbone forms seven hydrogen bonds with five CID side chains (Fig. 2a, b). The side chains of CTD residues Tyr 1b and Pro 3b are buried in the hydrophobic CID groove, and the phenolic hydroxyl group of Tyr 1b forms a hydrogen bond with Asp 68. This hydrogen bond explains why alanine replacement of CID residues 68–70 impairs CTD binding, and shows that mutation of Asp 68 to alanine can contribute to a temperature-sensitive yeast phenotype¹⁴. Substitution of Tyr 1 with phenylalanine in the CTD would result in the loss of this hydrogen bond, possibly contributing to the lethal effect of this mutation in yeast¹⁵. Furthermore, Tyr 1 phosphorylation, which has been observed in higher eukaryotes, is incompatible with the observed CTD–CID interaction. The CTD prolines Pro 3b and Pro 6b are both in *trans* conformation and their isomerization to a *cis* conformation would most probably impair CID binding.

The phosphate group of the CTD residue pSer 2b points away from the CID surface and does not contact it directly (Fig. 2), which is consistent with the ability of Pcf11 to bind to the unphosphorylated CTD¹¹. However, as phosphorylation of Ser 2 strongly enhances CTD binding^{11,14}, the CID domain may recognize the pSer 2b phosphate indirectly. This indirect pSer 2 recognition may result from stabilization of the CTD β -turn by an additional intramolecular hydrogen bond, formed between the pSer 2b phosphate and the side-chain hydroxyl group of Thr 4b (Fig. 2c, hydrogen bond C). Two more hydrogen bonds usually stabilize β -turns of the sequence SPXX¹⁶, and are observed in our structure (Fig. 2c, hydrogen bonds A and B). Formation of hydrogen bond B is impossible when Ser 2 is replaced by glutamate, a mutation that results in lethality in yeast¹⁵. The side chain of CTD residue Ser 5 also extends into solvent, and a phosphorylated Ser 5 residue cannot contact the CID domain, explaining why Ser 5 phosphorylation does not influence Pcf11 binding¹⁷.

Comparison of the CTD–CID complex structure with two published structures of CTD peptide–protein complexes^{18,19} highlights the diversity of CTD recognition modes and suggests that there is no simple structural basis for a CTD code. The structure of a Ser 2/Ser 5-phosphorylated CTD peptide bound to the WW domain of the prolyl isomerase Pin1, revealed an extended CTD conformation with the pSer 5 phosphate bound tightly and the pSer 2 phosphate exposed¹⁸. The structure of a Ser 5-phosphorylated CTD peptide bound to the Cgt1 subunit of the capping enzyme revealed contacts to pSer 5 phosphates and an extended CTD conformation¹⁹. In all three structures, the Pro 3 side chain docks into a hydrophobic pocket of the target protein, but binding of the Tyr 1 side chain occurs only in the CTD–CID and the CTD–Cgt1 complexes.

The CTD β -turn conformation observed in the CTD–CID complex can be pre-formed in the free CTD, as NMR studies of

CTD peptides in solution identified possible β -turns at the Ser2-Pro 3-Thr 4-Ser 5 and Ser 5-Pro 6-Ser 7-Tyr 1 motifs, with the first being more stable than the second^{20,21}. This indicates that the β -turn at the Ser 2-Pro 3-Thr 4-Ser 5 motif is most probably not the result of an

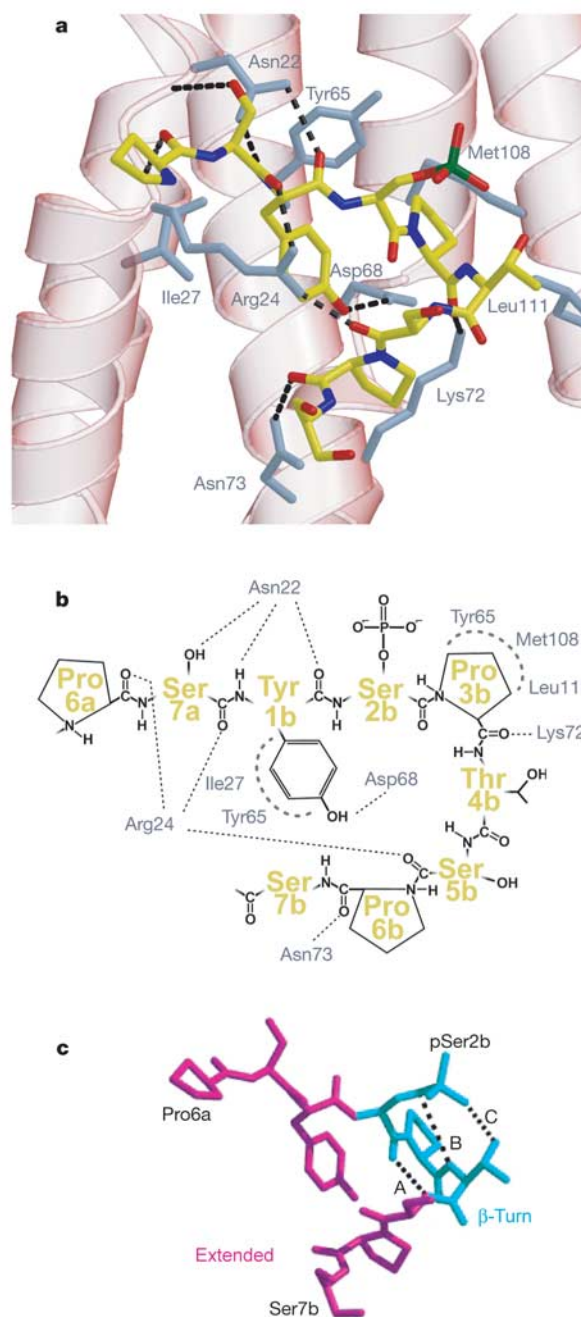


Figure 2 CTD–CID interactions. **a**, Binding of the CTD peptide to the CID domain. CID–CTD hydrogen bonds are indicated as broken black lines. CID residues involved in CTD interactions are depicted in grey. The peptide is coloured and oriented as in Fig. 1b–d. **b**, Schematic representation of observed CTD–CID interactions. Hydrophobic contacts with the side chains of Tyr 1b and Pro 3b are shown with curved broken grey lines and hydrogen bonds are indicated as broken black lines. **c**, CTD conformation and β -turn stabilization. CTD residues in an extended and β -turn conformation are shown in pink and cyan, respectively. Three turn-stabilizing hydrogen bonds are indicated with broken black lines (A, pSer 2b carbonyl–Ser 5b amide; B, pSer 2b γ -oxygen–Thr 4b amide; C, pSer 2b phosphate–Thr 4b hydroxyl). The view is as in **a**.

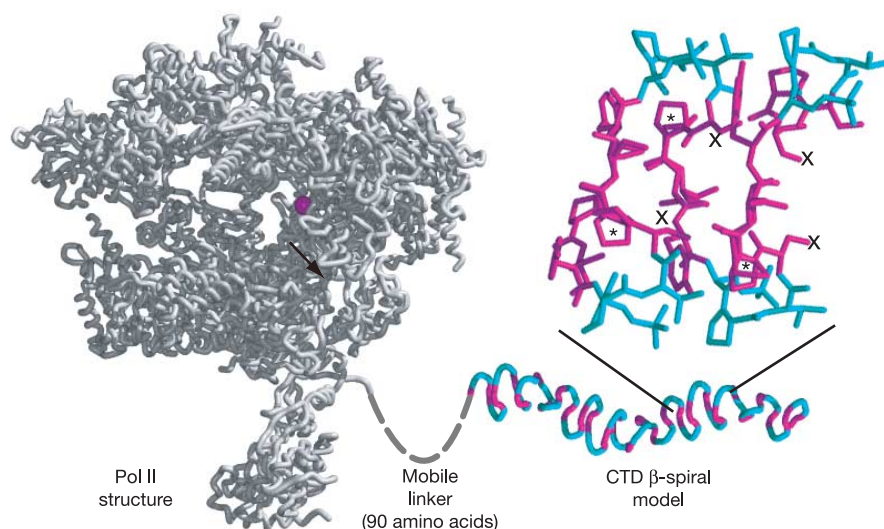


Figure 3 CTD β -spiral model. The complete 12-subunit Pol II structure³⁰ is shown in silver. A magenta sphere marks the active site. The arrow indicates the presumed direction of RNA exit. The CTD β -spiral model obtained from iterative superposition of the observed CTD peptide structure is shown as a coil, with alternating β -turns and extended regions in cyan and pink, respectively. The CTD is connected to the structured core of Pol

II with a mobile linker of approximately 90 amino acid residues in length (broken line). The yeast CTD model consists of 26 heptapeptide repeats and has a length of 100 Å. A detailed view of four CTD repeats is depicted above the spiral model. Asterisks indicate three superposed Pro 6-Ser 7 regions. Ser 5 side chains inside the spiral are labelled 'X'.

induced fit. Also arguing against formation of the β -turn after an induced fit is the observation that CID residue Arg 24, which binds before and within the β -turn (Fig. 2a, b), is not fully conserved (Fig. 1a), and is mobile in the free CID structure. By making the assumption that all CTD repeats adopt the observed conformation, we were able to model the entire CTD in one possible state. The backbone atoms of residues Pro 6a and Ser 7a of the CTD peptide structure were superimposed with those of Pro 6a and Ser 7b of a second peptide copy, which resulted in a model for subsequent CTD repeats with the Pro 6-Ser 7 regions overlapping (Fig. 3). In an iterative process we were able to model all 26 repeats of the yeast CTD without steric clashes or Ramachandran violations.

The resulting CTD model is a compact left-handed β -spiral with 1.8 CTD repeats per winding (Fig. 3). The spiral would only be 100 Å long, in contrast to a predicted length of 650 Å for a fully extended CTD model³. This may explain how the CTD can be accommodated in Pol II crystals³, and is largely consistent with a CTD density in electron micrographs of Pol II, although alternative compact CTD states may co-exist²². The compact β -spiral may represent one possible CTD conformation that can, however, be readily altered. Changes in the extended Ser 5-Pro 6-Ser 7-Tyr 1 region might invert the spiral-handedness, and formation of a second β -turn at this motif would create an open and extended β -spiral²¹. Local opening of the proposed spiral is required for access of CTD-modifying and -binding proteins, such as CID factors. Whereas Ser 2 phosphates are accommodated on the spiral surface, Ser 5 side chains are closely spaced inside the spiral, which is incompatible with phosphorylation. Thus, the Ser 5-phosphorylated CTD must be more extended, possibly explaining why the Stokes radius and the proteolytic sensitivity of the CTD both increase after phosphorylation²³.

We predict that a dynamic CTD β -spiral is involved in the mRNA transcription-processing cycle. The unphosphorylated CTD may largely form a compact β -spiral that is incorporated into the transcription initiation complex. Subsequent Ser 5 phosphorylation would open the spiral, resulting in an extended region that binds the capping enzyme. During elongation, Ser 2 phosphorylation stabil-

izes β -turns at motifs Ser 2-Pro 3-Thr 4-Ser 5, which triggers binding of CID factors for polyadenylation-dependent (Pcf11) and -independent (Nrd1) 3'-end formation, and may recreate compact spiral regions. We propose that during transcription termination, phosphatases access and remove the pSer 2 phosphate that is exposed in CTD-CID complexes, thereby releasing the CID factor. Dephosphorylation and release of all bound factors from the CTD would regenerate the initiation-competent form of Pol II and complete recycling of the enzyme. □

Methods

Sample preparation

The DNA encoding the *Saccharomyces cerevisiae* Pcf11 CID domain (residues 1–140) was amplified by polymerase chain reaction from yeast genomic DNA and was cloned into the pET28b expression vector (Novagen). This results in an open reading frame that includes three methionine residues and a C-terminal hexahistidine tag (variant CID). A fourth methionine was introduced by site-directed mutagenesis of residue L16 with the overlap extension method. The resulting variant CID-L16M was used for selenomethionine labelling in a standard procedure. CID was expressed overnight at 20 °C in *Escherichia coli* BL21 (DE3) CodonPlus RIL cells (Stratagene). Cells were collected, suspended in 50 mM Tris-HCl at pH 8.0, 300 mM NaCl and 10 mM β -mercaptoethanol and lysed with a French press. After incubation on ice with DNase I (Sigma) for 20 min, the slurry was cleared by centrifugation. Soluble lysates were subjected to chromatography with a Ni-NTA column (Qiagen), before separation on a MonoQ column (Amersham). Peak fractions were concentrated and purified on a Superose-12 size exclusion column (Amersham) equilibrated with 50 mM HEPES at pH 7.5, 100 mM NaCl, 1 mM EDTA and 1 mM dithiothreitol. Pure proteins were concentrated to 40 mg ml⁻¹. The CTD peptide was synthesized as a precursor with a C-terminal aminocaproic acid spacer.

Crystallization and data collection

Crystals of CID and CID-L16M were grown at 20 °C by the hanging-drop vapour diffusion method with a reservoir solution containing 25% ethylene glycol and 5% polyethylene glycol 550 monomethylester (PEG 550 MME). Crystals grew to a size of 0.3 × 0.15 × 0.15 mm within two days and were transferred to 35% ethylene glycol, 5% PEG 550 MME and flash-cooled in liquid nitrogen. For peptide soaks, crystals were stabilized in 35% ethylene glycol, 30% polyethylene glycol 3350. Peptide concentration and soaking times were optimized to prevent crystal cracking. Synchrotron diffraction data were collected at the European Synchrotron Radiation Facility (ESRF), Grenoble, and at the Swiss Light Source (SLS), Villigen. Data were complete to 2.1 and 2.2 Å resolution for the CID and the peptide-soaked CID crystals, respectively. Multiwavelength anomalous diffraction (MAD) data were collected from a selenomethionine-labelled CID-L16M crystal to 2.4 Å resolution. Data were processed

with HKL²⁴ or XDS²⁵. Crystals belong to space group $P2_12_12_1$ with unit cell dimensions $a = 56.7 \text{ \AA}$, $b = 67.7 \text{ \AA}$ and $c = 135.6 \text{ \AA}$, and contain three molecules per asymmetric unit.

Structure determination

Selenium sites were located with SnB²⁶. MAD phases were calculated with SHARP²⁷ and improved by density modification. A CID model was built with O²⁸ and refined with CNS²⁹. The refined model has excellent stereochemical quality and a free R -factor of 25.2%. A difference Fourier for the peptide-soaked crystal was calculated with phases from the CID model and identified strong density for the CTD peptide bound to one CID molecule (chain B). The two other CID molecules pack against each other, burying their peptide-binding sites. After peptide building, the CID–CTD model was refined to a free R -factor of 25.7%. Peptide binding does not result in significant conformational changes in the CID domain. Soaking experiments with phosphoserine did not reveal any additional electron density.

Received 4 March; accepted 21 May 2004; doi:10.1038/nature02679.

- Hirose, Y. & Manley, J. L. RNA polymerase II and the integration of nuclear events. *Genes Dev.* **14**, 1415–1429 (2000).
- Proudfoot, N. J., Furger, A. & Dye, M. J. Integrating mRNA processing with transcription. *Cell* **108**, 501–512 (2002).
- Cramer, P., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 Å resolution. *Science* **292**, 1863–1876 (2001).
- Komarnitsky, P., Cho, E. J. & Buratowski, S. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes Dev.* **14**, 2452–2460 (2000).
- Ahn, S. H., Kim, M. & Buratowski, S. Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3' end processing. *Mol. Cell* **13**, 67–76 (2004).
- Ni, Z., Schwartz, B. E., Werner, J., Suarez, J. R. & Lis, J. T. Coordination of transcription, RNA processing, and surveillance by P-TEFb kinase on heat shock genes. *Mol. Cell* **13**, 55–65 (2004).
- Buratowski, S. The CTD code. *Nature Struct. Biol.* **10**, 679–680 (2003).
- Yuryev, A. *et al.* The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins. *Proc. Natl Acad. Sci. USA* **93**, 6975–6980 (1996).
- Patturajan, M., Wei, X., Berezney, R. & Corden, J. L. A nuclear matrix protein interacts with the phosphorylated C-terminal domain of RNA polymerase II. *Mol. Cell. Biol.* **18**, 2406–2415 (1998).
- Steinmetz, E. J., Conrad, N. K., Brow, D. A. & Corden, J. L. RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts. *Nature* **413**, 327–331 (2001).
- Barilla, D., Lee, B. A. & Proudfoot, N. J. Cleavage/polyadenylation factor IA associates with the carboxyl-terminal domain of RNA polymerase II in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **98**, 445–450 (2001).
- Misra, S., Puertollano, R., Kato, Y., Bonifacio, J. S. & Hurley, J. H. Structural basis for acidic-cluster-dileucine sorting-signal recognition by VHS domains. *Nature* **415**, 933–937 (2002).
- Conti, E., Uy, M., Leighton, L., Blobel, G. & Kuriyan, J. Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin α . *Cell* **94**, 193–204 (1998).
- Sadowski, M., Dichtl, B., Hubner, W. & Keller, W. Independent functions of yeast Pcf11p in pre-mRNA 3' end processing and in transcription termination. *EMBO J.* **22**, 2167–2177 (2003).
- West, M. L. & Corden, J. L. Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations. *Genetics* **140**, 1223–1233 (1995).
- Suzuki, M. SPXX, a frequent sequence motif in gene regulatory proteins. *J. Mol. Biol.* **207**, 61–84 (1989).
- Licaltosi, D. *et al.* Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II. *Mol. Cell* **9**, 1101–1111 (2002).
- Verdecia, M. A., Bowman, M. E., Lu, K. P., Hunter, T. & Noel, J. P. Structural basis for phosphoserine-proline recognition by group IV WW domains. *Nature Struct. Biol.* **7**, 639–643 (2000).
- Fabrega, C., Shen, V., Shuman, S. & Lima, C. D. Structure of an mRNA capping enzyme bound to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Mol. Cell* **11**, 1549–1561 (2003).
- Kumaki, Y., Matsushima, N., Yoshida, H., Nitta, K. & Hikichi, K. Structure of the YSPSPS repeat containing two SPXX motifs in the CTD of RNA polymerase II: NMR studies of cyclic model peptides reveal that the SPSPS turn is more stable than SPSPS in water. *Biochim. Biophys. Acta* **1548**, 81–93 (2001).
- Cagas, P. M. & Corden, J. L. Structural studies of a synthetic peptide derived from the carboxyl-terminal domain of RNA polymerase II. *Proteins* **21**, 149–160 (1995).
- Meredith, G. D. *et al.* The C-terminal domain revealed in the structure of RNA polymerase II. *J. Mol. Biol.* **258**, 413–419 (1996).
- Zhang, J. & Corden, J. L. Phosphorylation causes a conformational change in the carboxyl-terminal domain of the mouse RNA polymerase II largest subunit. *J. Biol. Chem.* **266**, 2297–2302 (1991).
- Otwiniński, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
- Kabsch, W. Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J. Appl. Crystallogr.* **26**, 795–800 (1993).
- Smith, G. D., Nagar, B., Rini, J. M., Hauptman, H. A. & Blessing, R. H. The use of SnB to determine an anomalous scattering substructure. *Acta Crystallogr. D* **54**, 799–804 (1998).
- Terwilliger, T. C. Automated structure solution, density modification and model building. *Acta Crystallogr.* **58**, 1937–1940 (2002).
- Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Armache, K. J., Kettenberger, H. & Cramer, P. Architecture of initiation-competent 12-subunit RNA polymerase II. *Proc. Natl Acad. Sci. USA* **100**, 6964–6968 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements Part of this work was performed at the Swiss Light Source, Paul Scherrer Institut, Villigen, Switzerland. We thank C. Schulze-Briese and the staff of beamline X06SA for help; L. Jacquamet and J. McCarthy for help during data collection at the European Synchrotron Radiation Facility, Grenoble, France; G. Arnold for peptide synthesis; and K. Sträßer and members of the Cramer laboratory for careful reading of the manuscript. P.C. is supported by the Deutsche Forschungsgemeinschaft, the EMBO Young Investigator Programme and the Fonds der Chemischen Industrie. A.M. is supported by an EMBO long-term fellowship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to P.C. (cramer@lmb.uni-muenchen.de). Atomic coordinates and structure factors for the Pcf11 CID domain and the CTD–CID complex have been deposited in the Protein Data Bank under accession numbers 1SZ9 and 1SZA, respectively.

Cyclin-dependent kinases regulate the antiproliferative function of Smads

Isao Matsuura^{1,2,3}, Natalia G. Denissova^{1,2,3*}, Guannan Wang^{1,2,3*}, Dongming He^{1,2,3}, Jianyin Long^{1,2,3} & Fang Liu^{1,2,3}

¹Center for Advanced Biotechnology and Medicine, and ²Susan Lehman Cullman Laboratory for Cancer Research, Department of Chemical Biology, Ernest Mario School of Pharmacy, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA

³Cancer Institute of New Jersey, New Brunswick, New Jersey 08903, USA

* These authors contributed equally to this work

Transforming growth factor- β (TGF- β) potently inhibits cell cycle progression at the G1 phase^{1,2}. Smad3 has a key function in mediating the TGF- β growth-inhibitory response. Here we show that Smad3 is a major physiological substrate of the G1 cyclin-dependent kinases CDK4 and CDK2. Except for the retinoblastoma protein family^{3,4}, Smad3 is the only CDK4 substrate demonstrated so far. We have mapped CDK4 and CDK2 phosphorylation sites to Thr 8, Thr 178 and Ser 212 in Smad3. Mutation of the CDK phosphorylation sites increases Smad3 transcriptional activity, leading to higher expression of the CDK inhibitor p15. Mutation of the CDK phosphorylation sites of Smad3 also increases its ability to downregulate the expression of *c-myc*. Using Smad3^{-/-} mouse embryonic fibroblasts and other epithelial cell lines, we further show that Smad3 inhibits cell cycle progression from G1 to S phase and that mutation of the CDK phosphorylation sites in Smad3 increases this ability. Taken together, these findings indicate that CDK phosphorylation of Smad3 inhibits its transcriptional activity and antiproliferative function. Because cancer cells often contain high levels of CDK activity^{5,6}, diminishing Smad3 activity by CDK phosphorylation may contribute to tumorigenesis and TGF- β resistance in cancers.

Cell cycle progression from G1 to S phase is governed by CDK4 and the homologous CDK6 as well as CDK2 (ref. 4). CDK4 and CDK6 are activated by D-type cyclins at early to mid-G1 phase, whereas CDK2 is activated by E- and A-type cyclins during late G1 and S phase, respectively⁴. The activities of CDK4 and CDK6, and of CDK2, are constrained by the p16 INK4 family and the p21 Cip/Kip family inhibitors, respectively⁴.

TGF- β potently inhibits cell proliferation by causing cell cycle arrest at the G1 phase^{1,2}. Smad proteins can mediate TGF- β growth inhibitory responses^{1,2}. In the basal state, Smad2 and Smad3 are distributed throughout cells. In response to TGF- β , Smad2 and

Smad3 are phosphorylated at the carboxyl terminus by TGF- β receptor and form complexes with Smad4 (refs 1, 2). These complexes then accumulate in the nucleus and regulate the transcription of target genes that include cell cycle regulators^{1,2}, such as the CDK inhibitors p15 and p21 and the protooncogene *c-myc* (refs 7–16). Smad3 is important in antiproliferative responses. For instance, hyperproliferation is a component of the carcinogenic process that leads to the development of metastatic colon cancer in Smad3^{-/-} mice¹⁷. The loss of Smad3 expression increases susceptibility to tumorigenicity in human gastric cancer¹⁸. Mice lacking Smad3 display squamous hyperplasia in the stomach (C.-X. Deng, personal communication). Smad3^{-/-} mice also show accelerated wound healing, partly owing to an increased rate of re-epithelialization¹⁹. A variety of primary cells from Smad3^{-/-} mice are resistant to the growth inhibitory effects of TGF- β (refs 20–22), indicating that

Smad3 has a key function in responsiveness to TGF- β .

Because Smads contain potential CDK phosphorylation sites, we analysed whether they are substrates for CDKs. To investigate whether cyclin D–CDK4 can phosphorylate Smads, we performed an *in vitro* kinase assay with endogenous CDK4 immunoprecipitated from Mv1Lu mink lung epithelial cells by an affinity-purified CDK4-specific antibody. Full-length Smad proteins fused to glutathione S-transferase (GST) were used as substrates. Two frequently used fusion proteins of retinoblastoma protein (Rb) with GST were included as positive controls. As shown in Fig. 1a, the highly homologous Smad2 and Smad3 were phosphorylated. Smad4, which is homologous to Smad2 and Smad3 in the amino-terminal and C-terminal domains but divergent from Smad2 and Smad3 in the middle proline-rich region, was not phosphorylated (Fig. 1a), nor was GST alone (data not shown). The phosphorylation was

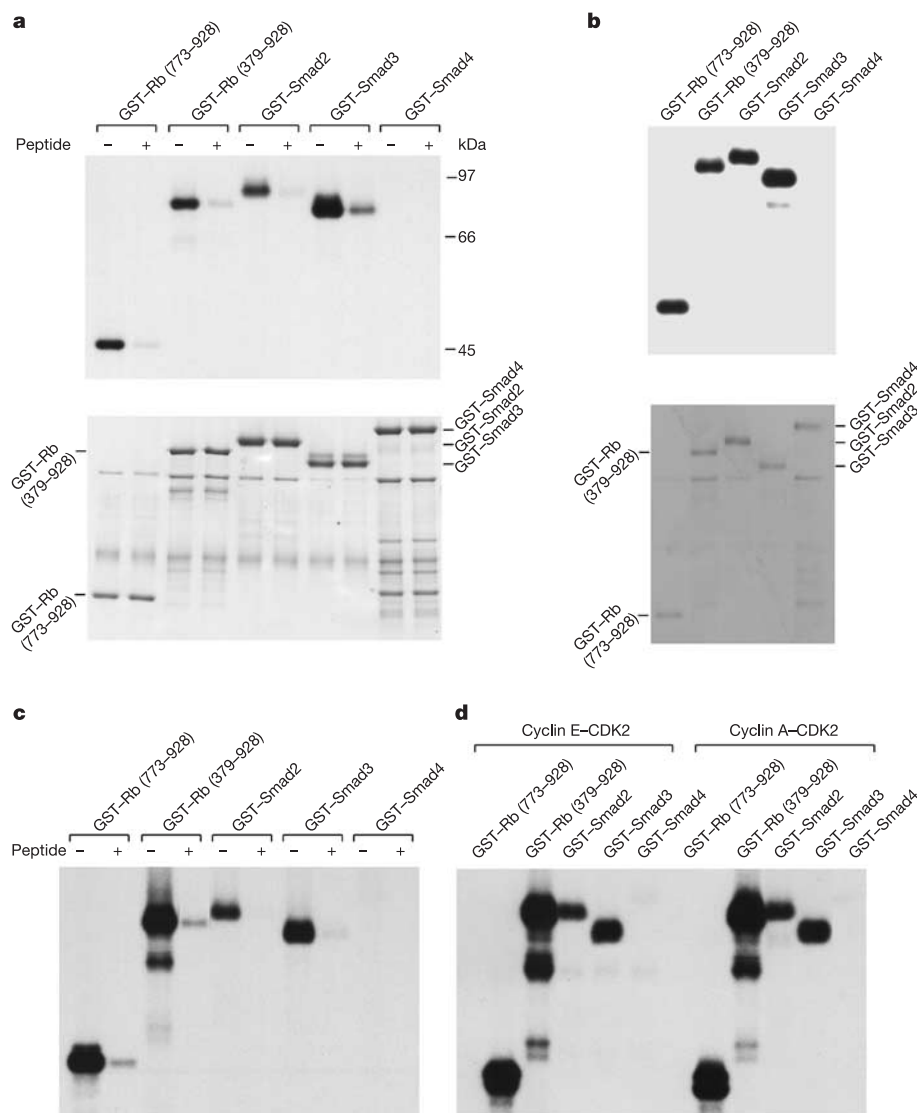


Figure 1 CDK4 and CDK2 can phosphorylate Smad3 and Smad2 *in vitro*. **a**, Top: immunoprecipitated (IP) CDK4 kinase assay. CDK4 immunoprecipitated from 240 μ g Mv1Lu cell lysate with 1.2 μ g CDK4 antibody in the absence or presence of 9 μ g of the antigen peptide (mouse CDK4 amino acids 282–303) was used in a kinase assay with 1 μ M substrates. Bottom: protein amounts indicated by Coomassie blue staining. **b**, Top: reconstituted CDK4 kinase assay. Reconstituted cyclin D–CDK4 (500 ng) was used to phosphorylate 0.4 μ M substrates. Bottom: protein level indicated by Coomassie blue

staining. **c**, Immunoprecipitated CDK2 kinase assay. CDK2 immunoprecipitated from 240 μ g Mv1Lu cell lysate with 1.2 μ g CDK2 antibody in the absence or presence of 6 μ g of the antigen peptide (human CDK2 amino acids 283–298) was used in a kinase assay with 1 μ M substrates. **d**, Reconstituted CDK2 kinase assay. Reconstituted cyclin E–CDK2 (13 ng) and reconstituted cyclin A–CDK2 (96 ng) were used to phosphorylate 1 μ M substrates.

specific, because preincubation of the antibody with the CDK4 antigen peptide before immunoprecipitation significantly inhibited phosphorylation (Fig. 1a). Because CDK4 has a very strict substrate specificity, indicated by its inability to phosphorylate the canonical CDK substrate histone H1 (ref. 4), we further verified that CDK4 can indeed phosphorylate Smads by using bacterially expressed and *in vitro* reconstituted CDK4 (Fig. 1b). Notably, Smad3 was phosphorylated to a greater extent than Rb by immunoprecipitated CDK4 from Mv1Lu cells (Fig. 1a) and many other cell lines (data not shown) as well as by reconstituted CDK4 (Fig. 1b). Further substrate titration experiments comparing Smad3 and Rb phosphorylation by using either immunoprecipitated CDK4 or reconstituted CDK4 confirmed that Smad3 is an excellent substrate for CDK4 (Supplementary Fig. S1 and Supplementary Table S1). Similar experiments indicated that Smad3 and Smad2, but not

Smad4, can also be specifically phosphorylated by immunoprecipitated CDK2 (Fig. 1c) and bacterially expressed and *in vitro* reconstituted cyclin E-CDK2 or cyclin A-CDK2 complexes (Fig. 1d).

To determine whether endogenous Smad3 is phosphorylated by G1 CDKs, we synchronized Mv1Lu cells at the G0/G1 phase by contact inhibition as described previously²³. Cells were then released from growth arrest by plating into fresh medium and, at different time points, were labelled with ³²P-orthophosphate and immunoprecipitated by a Smad3-specific antibody to analyse the endogenous Smad3 phosphorylation status. Unlabelled cells prepared in parallel were analysed by flow cytometry to determine the cell cycle distribution at each time point. As shown in Fig. 2a, Smad3 phosphorylation oscillated in a cell-cycle-dependent manner. The maximal phosphorylation of Smad3 occurred at the G1/S

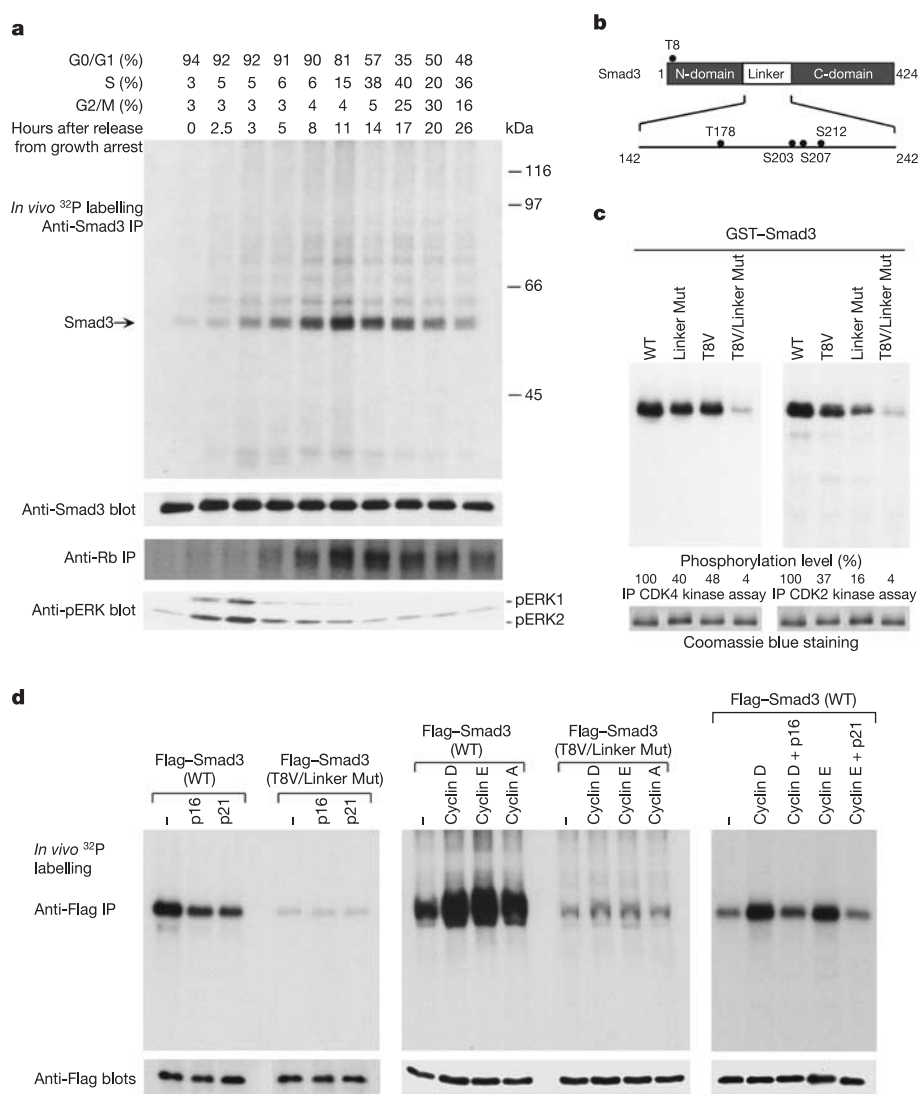


Figure 2 Smad3 is phosphorylated by endogenous G1 CDKs *in vivo*. **a**, Endogenous Smad3 is phosphorylated in a cell cycle-dependent manner. Mv1Lu cells were synchronized by contact inhibition, then plated into fresh medium, labelled with ³²P-orthophosphate, immunoprecipitated (IP) first with a Smad3 antibody and subsequently with an anti-Rb antibody. The ERK activity profile, analysed by a phosphotyrosine antibody that recognizes only activated ERK (pERK1 and pERK2) in unlabelled lysates, indicates that ERK might contribute to Smad3 phosphorylation at very

early, but not at the peak, phosphorylation time points. **b**, Schematic diagram of Smad3 structure. **c**, Mutation of the Thr 8 and four phosphorylation sites in the Smad3 linker region markedly decreases CDK4 and CDK2 phosphorylation *in vitro*. WT, wild-type. **d**, Left, p16 or p21 transfected into Mv1Lu/L17 cells can inhibit endogenous CDK to phosphorylate Smad3. Middle and right, cyclins transfected into COS cells can activate endogenous CDKs to phosphorylate Smad3, and p16 and p21 can inhibit this activity.

junction, indicating that Smad3 was phosphorylated by G1 CDKs. Subsequent immunoprecipitation of the ^{32}P -labelled cell lysates with an antibody against Rb showed that the peak of Rb phosphorylation slightly lagged behind that of Smad3 phosphorylation (Fig. 2a), suggesting that Smad3 is a good substrate for CDK4 *in vivo*.

Smad3 contains nine potential CDK phosphorylation sites, four of which are located in the proline-rich linker region: Thr 178, Ser 203, Ser 207 and Ser 212 (Fig. 2b and Supplementary Fig. S2). The threonine at amino acid position 8 is potentially a good site, particularly for CDK4 (ref. 24). Through a number of mutational studies, we found that simultaneous mutation of Thr 8 and the four sites in the linker, designated Smad3 (T8V/Linker Mut), markedly decreased phosphorylation by CDK4 or CDK2 (Fig. 2c), indicating that CDK4 and CDK2 phosphorylation might occur within these

five sites *in vitro*.

In transient transfection assays, Flag-Smad3 was phosphorylated, and p16 and p21 each inhibited the phosphorylation (Fig. 2d, left panel). Smad3 (T8V/Linker Mut) was phosphorylated to a much lower level than the wild-type Smad3. To determine whether the introduction of exogenous cyclins can activate endogenous CDK to phosphorylate Smad3, the wild-type or mutant version of Flag-Smad3 was cotransfected either individually or together with cyclins D, E or A. As shown in Fig. 2d (middle panel), the phosphorylation of wild-type Flag-Smad3 was significantly increased by cotransfected cyclins. The effects of cyclins D or E can be inhibited by cotransfected p16 or p21, respectively (Fig. 2d, right panel). In contrast, phosphorylation of Smad3 (T8V/Linker Mut) was only slightly increased by cyclins D or E. These observations provide additional evidence that Smad3 is phosphorylated by CDK4 and CDK2 *in vivo* and that the phosphorylation occurs within these five sites.

To determine the exact CDK phosphorylation sites in Smad3, we generated phosphopeptide antibodies against each of the five potential phosphorylation sites: Thr 8 and the four sites in the linker (Thr 178, Ser 203, Ser 207 and Ser 212). Each of the five sites was phosphorylated by both CDK4 and CDK2 *in vitro*, and only Thr 8, Thr 178 and Ser 212 were phosphorylated by CDK4 and CDK2 *in vivo* (Supplementary Figures S3–S5). To confirm that the other four potential sites (Thr 131, Ser 316, Ser 391 and Ser 415) cannot be phosphorylated by CDK, we also generated phosphopeptide antibodies against each of these sites and found that these four sites indeed cannot be phosphorylated by CDK4 or CDK2 (data not shown). Mitogen-activated protein (MAP) kinase was shown to phosphorylate Smad3 in the linker region²⁵. We found that Ser 203 and Ser 207 were phosphorylated by MAP kinase and that Thr 178 was phosphorylated mostly by CDK and to a lesser extent by MAP kinase (Supplementary Fig. S5).

To analyse the role of CDK phosphorylation of Smad3, we examined the effect of mutation of the Smad3 CDK phosphorylation sites on the p15 reporter gene. As shown in Fig. 3a, each of T8V, T178V and S212A has a higher activity than the wild-type Smad3 to stimulate the p15 promoter, and the triple mutant (T8V/T178V/S212A) has the highest activity. Moreover, the GAL4-Smad3 triple mutant (T8V/T178V/S212A) has a higher activity than the wild-type GAL4-Smad3 on a GAL4-luciferase reporter gene in the absence as well as in the presence of different concentrations of TGF- β (Fig. 3b). We also found that cotransfection of CDK4 RNA-mediated interference (RNAi) or CDK2 RNAi constructs increases the basal and TGF- β -induced p15 reporter gene activity (Fig. 3c). Taken together, these results indicate that CDK phosphorylation of Smad3 can inhibit its transcriptional activity.

Smad3 also plays a critical role in the downregulation of *c-myc* expression by TGF- β (refs 13–16), which is necessary for subsequent p15 and p21 induction¹. Mutation of Smad3 CDK phosphorylation sites also increased its ability to downregulate *c-myc* (Fig. 3d). Downregulation of *c-myc* might involve an active repression mechanism, a possibility that requires further investigation.

The above observations prompted us to ask whether Smad3 can inhibit cell cycle progression from G1 to S phase, and whether mutation of the CDK phosphorylation sites renders it more effective in executing this function. Previous studies have shown that Smad3 together with Smad2 and Smad4 activates p15 expression, and introduction of Smad3 into Smad3^{-/-} mouse embryonic fibroblasts (MEF) restores the p15 reporter gene induction by TGF- β (ref. 11). Smad3^{-/-} primary MEF proliferated faster than the wild-type littermate MEF in the ³H-thymidine incorporation assay, and the TGF- β growth-inhibitory effects were largely lost in the Smad3^{-/-} primary MEF (Fig. 4a and ref. 21). To determine whether mutation of the CDK phosphorylation sites enables Smad3 to be more effective at inhibiting G1 cell cycle progression, we generated retroviral vectors that express either the wild-type Smad3 or the

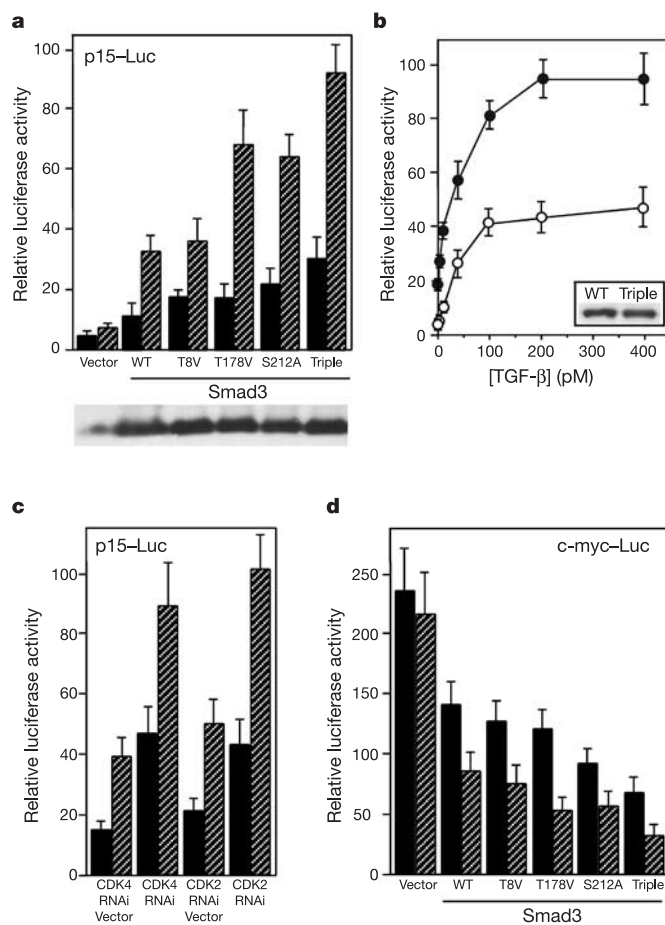


Figure 3 Mutation of CDK phosphorylation sites in Smad3 leads to an increased p15 level and reduced *c-myc* expression in reporter gene assays. **a**, Smad3^{-/-} MEF were transfected with a p15 reporter gene together with the wild-type or various CDK phosphorylation mutants of Smad3. Similar results were obtained in HepG2 cells, and HepG2 cells were used to examine the Smad3 protein expression levels, as shown in the gel at the bottom. Black bars, without TGF- β ; hatched bars, with TGF- β . **b**, Mv1Lu/L17 cells were transfected with a GAL4 reporter gene and TGF- β receptor along with either GAL4-Smad3 (open circles; WT) or GAL4-Smad3 (filled circles; Triple Mut), which contains the T8V, T178V and S212A mutations. **c**, CDK4 RNAi or CDK2 RNAi increases p15 reporter gene activity as analysed in HepG2 cells. Black bars, without TGF- β ; hatched bars, with TGF- β . **d**, Smad3^{-/-} MEF were transfected with a *c-myc* reporter gene along with the wild-type or various CDK phosphorylation mutants of Smad3 and analysed as indicated. Black bars, without TGF- β ; hatched bars, with TGF- β . Error bars show standard deviation (s.d.) of at least three independent transfection results.

various CDK phosphorylation mutants of Smad3, and then used them to infect Smad3^{-/-} primary MEF. As shown in Fig. 4b, wild-type Smad3 increased the cell population in G0/G1 phase and decreased the cell population in S phase. The various CDK phosphorylation mutants of Smad3 augmented these effects. Accord-

ingly, these mutants were more effective than the wild-type Smad3 in inhibiting cell proliferation as measured by the ³H-thymidine incorporation assay (Fig. 4c) and cell number (Fig. 4d), accompanied by increased p15 expression and lower c-Myc levels (Fig. 4e). In addition to the Smad3^{-/-} primary MEF, we have also

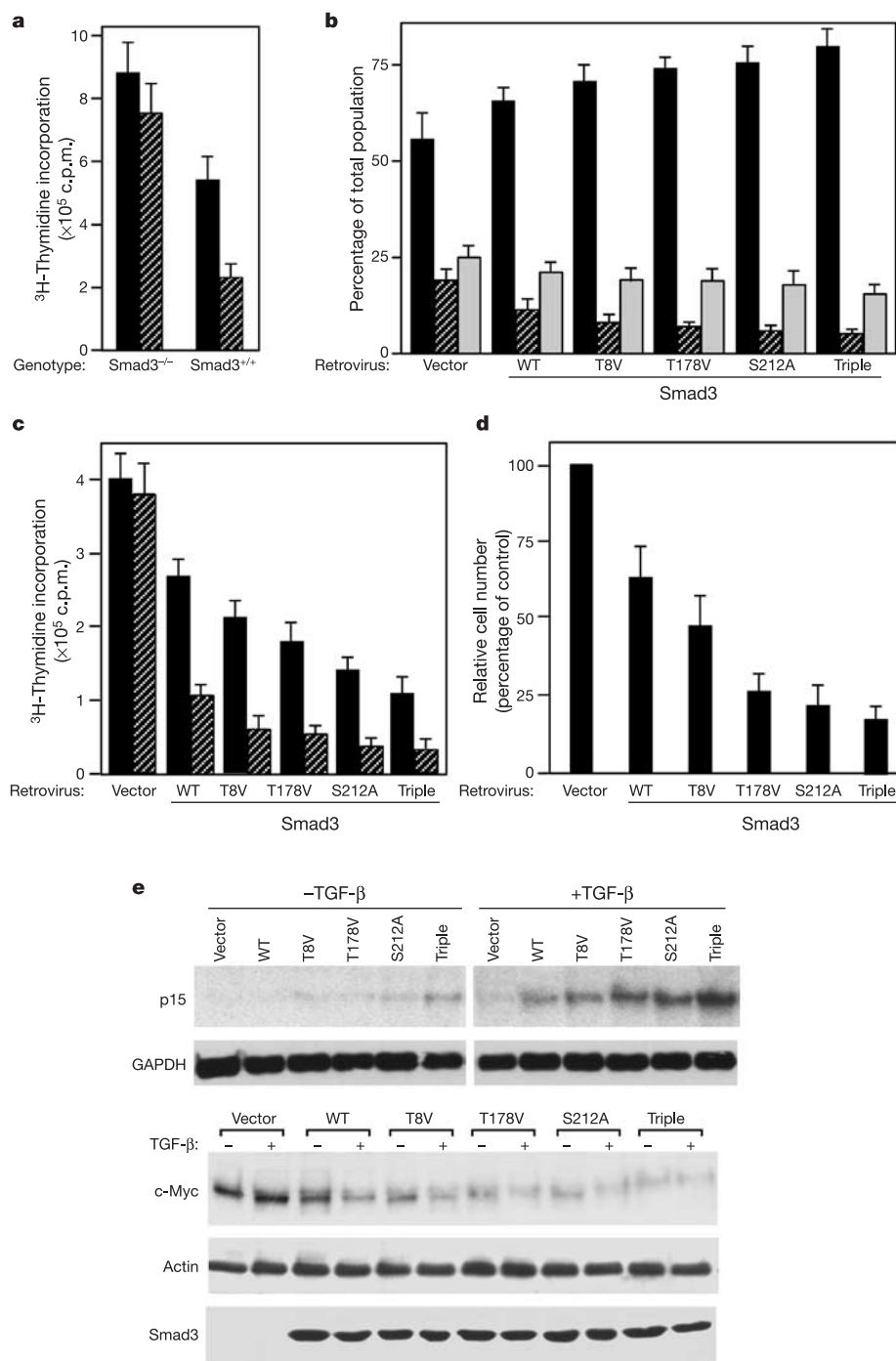


Figure 4 Mutation of the CDK phosphorylation sites in Smad3 leads to increased antiproliferative activities. **a**, Smad3^{-/-} primary MEF and the wild-type littermate control MEF (both at passage 3) were compared in a ³H-thymidine incorporation assay in the absence (black bars) and in the presence (hatched bars) of TGF-β. The average of four experiments is shown. Error bars represent standard deviation (s.d.). **b–e**, Smad3^{-/-} primary MEF (passage 3) were infected with retrovirus vector, retroviral wild-type Smad3 or various Smad3 CDK phosphorylation mutants. Infected cells were then split and seeded

for fluorescence-activated cell sorting analysis 2 days later (**b**; black bars, G0/G1 phase; hatched bars, S phase; grey bars, G2/M phase), ³H-thymidine incorporation assay (**c**; black bars, without TGF-β; hatched bars, with TGF-β), measurement of cell number about 5 days later (**d**) and analysis of p15 and c-Myc levels (**e**). The error bars in **b–d** represent s.d. of four experiments. The infected MEF were treated with TGF-β for 24 h in **e**. Total RNA (10 μg) and protein (15 μg) were analysed by northern blot (top) and immunoblotting (bottom), respectively.

observed the same trend in other cell types including the Mv1Lu epithelial cells, which contain relatively low levels of wild-type Smad3 (data not shown). Taken together, these observations indicate that phosphorylation of Smad3 by CDK facilitates cell cycle progression from G1 to S phase.

Thus, we have shown that Smad3 is phosphorylated by CDK4 and CDK2. Mutation of its CDK phosphorylation sites increases its transcriptional activity and antiproliferative function. We propose that under physiological conditions, phosphorylation of Smad3 by CDK inhibits its transcriptional activity, contributing to a decreased level of p15 and an increased level of c-Myc, thus facilitating cell cycle progression from G1 to S phase. In the presence of physiological concentrations of TGF- β , normal cells are inhibited by TGF- β . However, cancer cells often contain high levels of CDK activities because of frequent amplification, translocation or overexpression of the cyclin D1 gene or inactivation of the tumour suppressor p16 (refs 5, 6). In addition, overexpression of cyclin E and decreases in CDK inhibitor p27 levels also occur in cancer cells²⁶. Inactivation of Smad3 and presumably the homologous Smad2 by extensive CDK phosphorylation may provide an important mechanism for resistance to the TGF- β growth-inhibitory effects in cancers. □

Methods

Phosphorylation *in vitro* by immunoprecipitated CDK

Mv1Lu cells were lysed in a buffer containing 50 mM Tris pH 7.5, 150 mM NaCl, 0.5% Nonidet P40, 1 mM dithiothreitol (DTT) and protease and phosphatase inhibitors. Antibodies against CDK2 (M2) and CDK4 (C-22) from Santa Cruz Biotechnology were used for immunoprecipitations. The kinase reaction was carried out for 1 h in 30 μ l containing 50 mM HEPES pH 7.4, 15 mM MgCl₂, 1 mM EGTA, 0.1% Tween 20, 1 mM DTT, 50 μ M ATP, 5 μ Ci [γ -³²P]ATP (3000 Ci mmol⁻¹) and substrates at 30 °C. GST-Rb (773–928) contains the proline-rich region, and GST-Rb (379–928) contains in addition the Rb pocket domain.

Phosphorylation *in vitro* by reconstituted CDK

GST-CDK4, GST-CDK2 and His-tagged cyclins D, E and A were expressed and purified from *Escherichia coli* as described previously²⁷. GST-CDK and His-cyclin were mixed and incubated with HeLa extracts in the presence of ATP and Mg²⁺ to activate CDK4 or CDK2. Cyclin D-CDK4 was reconstituted as described previously²⁸ except that no MnCl₂ was included. The cyclin D-CDK4 complex was then purified with glutathione agarose beads. Cyclin E-CDK2 and cyclin A-CDK2 were reconstituted as described previously²⁷ and purified by successive Ni-nitrilotriacetate-agarose and glutathione-agarose chromatography. The purified cyclin-CDK complexes were used to phosphorylate substrates for 40 min in 20 μ l reactions containing 35 mM HEPES pH 7.4, 10 mM MgCl₂, 1 mM EGTA, 0.1% Tween 20, 1 mM DTT, 15 μ M ATP and 5 μ Ci [γ -³²P]ATP for CDK4 or 100 μ M ATP and 2 μ Ci [γ -³²P]ATP for CDK2 at 30 °C.

Phosphorylation *in vivo*

For analysis of endogenous Smad3 phosphorylation, Mv1Lu mink lung epithelial cells were synchronized at G0/G1 phase by contact inhibition in complete medium as previously described²³. In brief, Mv1Lu cells were grown to full confluence in complete medium. Cells were then split and plated into fresh medium. At different time points, cells were phosphate-starved for 0.5 h and then labelled for 1.5 h with 1 mCi ml⁻¹ ³²P-orthophosphate. Cells were lysed in a buffer containing 10 mM Tris pH 7.8, 150 mM NaCl, 1 mM EDTA, 1% Nonidet P40, and protease and phosphatase inhibitors, and immunoprecipitated by a Smad3-specific antibody. For analysis of transfected Flag-tagged Smad3 phosphorylation, Mv1Lu/L17 or COS cells were transfected by DEAE-dextran. At 30–36 h after transfection, cells were phosphate-starved for 45 min and labelled with ³²P-orthophosphate at 1 mCi ml⁻¹ for 2.5 h followed by immunoprecipitation by a Flag antibody.

Retrovirus production

Wild-type or CDK phosphorylation mutant Smad3 were subcloned into the retroviral vector pLZRSΔ-IRES-GFP²⁹ and transfected into ecotropic phoenix packaging cells to produce retroviruses as described previously³⁰. Smad3^{-/-} MEF were infected at greater than 90% efficiency.

Generation of MEF and ³H-thymidine incorporation assay

Smad3^{+/-} mice were crossed and MEF were generated as described previously²¹; 2 × 10⁵ Smad3^{-/-} primary MEF and the wild-type littermate control MEF (both at passage 3) were seeded in six-well plates for 24 h, then treated for 24 h with or without 500 pM TGF- β . During the last 4 h, 5 μ Ci of ³H-thymidine was added to the culture, and ³H-thymidine incorporation was assayed as described previously²¹.

Received 5 January; accepted 10 May 2004; doi:10.1038/nature02650.

1. Massagué, J., Blain, S. W. & Lo, R. S. TGF- β signaling in growth control, cancer, and heritable

disorders. *Cell* **103**, 295–309 (2000).

2. Derynck, R., Akhurst, R. J. & Balmain, A. TGF- β signaling in tumor suppression and cancer progression. *Nature Genet.* **29**, 117–129 (2001).
3. Weinberg, R. A. The retinoblastoma protein and cell cycle control. *Cell* **81**, 323–330 (1995).
4. Sherr, C. J. & Roberts, J. M. CDK inhibitors: positive and negative regulators of G1-phase progression. *Genes Dev.* **13**, 1501–1512 (1999).
5. Hall, M. & Peters, G. Genetic alterations of cyclins, cyclin-dependent kinases, and Cdk inhibitors in human cancer. *Adv. Cancer Res.* **68**, 67–108 (1996).
6. Sherr, C. J. Cancer cell cycles. *Science* **274**, 1672–1677 (1996).
7. Hannon, G. J. & Beach, D. p15^{INK4B} is a potential effector of TGF- β -induced cell cycle arrest. *Nature* **371**, 257–261 (1994).
8. Li, J. M., Nichols, M. A., Chandrasekharan, S., Xiong, Y. & Wang, X.-F. Transforming growth factor β activates the promoter of cyclin-dependent kinase inhibitor p15^{INK4B} through an Sp1 consensus site. *J. Biol. Chem.* **270**, 26750–26753 (1995).
9. Moustakas, A. & Kardassis, D. Regulation of the human p21/WAF1/Cip1 promoter in hepatic cells by functional interaction between Sp1 and Smad family members. *Proc. Natl Acad. Sci. USA* **95**, 6733–6738 (1998).
10. Pardali, K. *et al.* Role of Smad proteins and transcription factor Sp1 in p21^{Waf1/Cip1} regulation by transforming growth factor- β . *J. Biol. Chem.* **275**, 29244–29256 (2000).
11. Feng, X.-H., Lin, X. & Derynck, R. Smad2, Smad3 and Smad4 cooperate with Sp1 to induce p15^{INK4B} transcription in response to TGF- β . *EMBO J.* **19**, 5178–5193 (2000).
12. Seoane, J., Le, H. V., Shen, L., Anderson, S. A. & Massagué, J. Integration of Smad and forkhead pathways in the control of neuroepithelial and glioblastoma cell proliferation. *Cell* **117**, 211–223 (2004).
13. Chen, C. R., Kang, Y. & Massagué, J. Defective repression of c-myc in breast cancer cells: A loss at the core of the transforming growth factor growth arrest program. *Proc. Natl Acad. Sci. USA* **98**, 992–999 (2001).
14. Yagi, K. *et al.* c-myc is a downstream target of the Smad pathway. *J. Biol. Chem.* **277**, 854–861 (2002).
15. Chen, C.-R., Kang, Y., Siegel, P. M. & Massagué, J. E2F4/5 and p107 as Smad cofactors linking the TGF- β receptor to c-myc repression. *Cell* **110**, 19–32 (2002).
16. Frederick, J. P., Liberati, N. T., Waddell, D. S., Shi, Y. & Wang, X. F. Transforming growth factor beta-mediated transcriptional repression of c-myc is dependent on direct binding of Smad3 to a novel repressive Smad binding element. *Mol. Cell. Biol.* **24**, 2546–2559 (2004).
17. Zhu, Y., Richardson, J. A., Parada, L. F. & Graff, J. M. Smad3 mutant mice develop metastatic colorectal cancer. *Cell* **94**, 703–714 (1998).
18. Han, S. U. *et al.* Loss of the Smad3 expression increases susceptibility to tumorigenicity in human gastric cancer. *Oncogene* **23**, 1333–1341 (2004).
19. Ashcroft, G. S. *et al.* Mice lacking Smad3 show accelerated wound healing and an impaired local inflammatory response. *Nature Cell Biol.* **1**, 260–266 (1999).
20. Yang, X. *et al.* Targeted disruption of SMAD3 results in impaired mucosal immunity and diminished T cell responsiveness to TGF- β . *EMBO J.* **18**, 1280–1291 (1999).
21. Datto, M. B. *et al.* Targeted disruption of Smad3 reveals an essential role in transforming growth factor β -mediated signal transduction. *Mol. Cell. Biol.* **19**, 2495–2504 (1999).
22. Rich, J. N., Zhang, M., Datto, M. B., Bigner, D. D. & Wang, X.-F. Transforming growth factor- β -mediated p15^{INK4B} induction and growth inhibition in astrocytes is Smad3-dependent and a pathway prominently altered in human glioma cell lines. *J. Biol. Chem.* **274**, 35053–35058 (1999).
23. Laiho, M., DeCaprio, J. A., Ludlow, J. W., Livingston, D. M. & Massagué, J. Growth inhibition by TGF- β linked to suppression of retinoblastoma protein phosphorylation. *Cell* **62**, 175–185 (1990).
24. Kitagawa, M. *et al.* The consensus motif for phosphorylation by cyclin D1-Cdk4 is different from that for phosphorylation by cyclin A/E-Cdk2. *EMBO J.* **15**, 7060–7069 (1996).
25. Kretschmar, M., Doody, J., Timokhina, I. & Massagué, J. A mechanism of repression of TGF- β /Smad signaling by oncogenic ras. *Genes Dev.* **13**, 804–816 (1999).
26. Porter, P. L. *et al.* Expression of cell-cycle regulators p27Kip1 and cyclin E, alone and in combination, correlate with survival in young breast cancer patients. *Nature Med.* **3**, 222–225 (1997).
27. Matsuura, I. & Wang, J. Demonstration of cyclin-dependent kinase inhibitory serine/threonine kinase in bovine thymus. *J. Biol. Chem.* **271**, 5443–5450 (1996).
28. Phelps, D. E. & Xiong, Y. Assay for activity of mammalian cyclin D-dependent kinases CDK4 and CDK6. *Methods Enzymol.* **283**, 194–204 (1997).
29. Kim, M. J. *et al.* Nkx3.1 mutant mice recapitulate early stages of prostate carcinogenesis. *Cancer Res.* **62**, 2999–3004 (2002).
30. Swift, S., Lorens, J., Achacoso, P. & Nolan, G. P. Rapid production of retroviruses for efficient gene delivery to mammalian cells using 293T cell-based systems. In *Curr. Protocols Immunol.* (eds Coligan, J. E. *et al.*) Unit 10.28, Suppl. 31 (John Wiley & Sons, 1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are very grateful to X.-F. Wang for Smad3 mutant mice; E. P. Reddy, R. V. Mettus and H. Kiyokawa for CDK4 mutant mice; J. Massagué for reagents and continued support; C.-X. Deng for reagents and communications before publication; M. M. Shen and members of his laboratory for assistance with mouse work; C. Abate-Shen, M. Cobb, X.-F. Feng, J. Germino, G. J. Hannon, S.-J. Kim, M. Kretschmar, H. Lee, M.-H. Lee, E. Lees, X. Liu, H. L. Moses, G. Nolan, A. Rabson, D. Reinberg, M. Reiss, Y. Shi, N. Walworth and W. Xie for reagents and/or suggestions; and numerous colleagues for discussions. This work was supported by the 1999 American Association for Cancer Research–National Foundation for Cancer Research Career Development Award, a Burroughs Wellcome Fund New Investigator Award, a Kimmel Scholar Award from the Sidney Kimmel Foundation for Cancer Research, the Emerald Foundation, the New Jersey Commission on Cancer Research, and the NIH (to F.L.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to F.L. (fangliu@cabm.rutgers.edu).

A membrane-access mechanism of ion channel inhibition by voltage sensor toxins from spider venom

Seok-Yong Lee & Roderick MacKinnon

The Howard Hughes Medical Institute and Rockefeller University, 1230 York Avenue, New York, New York 10021, USA

Venomous animals produce small protein toxins that inhibit ion channels with high affinity. In several well-studied cases the inhibitory proteins are water-soluble and bind at a channel's aqueous-exposed extracellular surface^{1–4}. Here we show that a voltage-sensor toxin (VSTX1) from the Chilean Rose Tarantula (*Grammostola spatulata*) reaches its target by partitioning into the lipid membrane. Lipid membrane partitioning serves two purposes: to localize the toxin in the membrane where the voltage sensor resides and to exploit the free energy of partitioning to achieve apparent high-affinity inhibition. VSTX1, small hydrophobic poisons and anaesthetic molecules reveal a common theme of voltage sensor inhibition through lipid membrane access. The apparent requirement for such access is consistent with the recent proposal that the sensor in voltage-dependent K⁺ channels is located at the membrane–protein interface^{5,6}.

Many protein toxins from the venoms of scorpions, snakes and bees inhibit K⁺ channels by attacking the pore^{1–4}. These toxins bind to the extracellular water-exposed surface with high affinity (K_d in the nM range) and typically with very high association rate constants ($>10^7 \text{ M}^{-1} \text{ s}^{-1}$) (refs 7, 8). As one might anticipate for such high-affinity protein–protein interactions, the binding site on the channel for these toxins is reasonably large, encompassing at least 400 Å² (refs 2, 9, 10).

Tarantula venom toxins alter channel gating by binding to the voltage sensor of voltage-dependent ion channels^{11,12} (V. Ruta and R. M., submitted). They inhibit with apparent high affinity (effective inhibitory concentration in the 10 nM to 100 nM range)^{13,14}, probably bind to the channel with a 4-to-1 stoichiometry (one toxin molecule per subunit)¹¹, and contact amino acids within the voltage sensor^{12,15}. From the crystal structure of a voltage-dependent K⁺ channel, KvAP, we see that toxins bind to amino acids located in a helical segment called S3b, which along with S4 forms the helix–turn–helix ‘voltage sensor paddle’^{6,12}. Two properties of voltage-sensor toxins have been enigmatic. They have very small apparent association and dissociation rate constants, 10³ to 10⁴ times slower than the pore blocking toxins^{13,14}, and their binding site on the channel, consisting of only a few amino acids on S3b, seems too small to account for the apparent high-affinity inhibition^{10,15,16}.

The structures of several voltage sensor toxins have been determined^{17–19}. Typical of small disulphide core proteins from venoms, they are well-structured globular proteins despite their small size of about 35 amino acids. The voltage sensor toxins exhibit the unique feature of a strong hydrophobic moment, with one face being almost exclusively hydrophobic. Scanning mutagenesis of SGTx1 (the toxin from the spider *Scodra griseipes*) showed that the hydrophobic face is important for apparent affinity²⁰. To test whether hanatoxin (from *G. spatulata*) might bind to membranes, octanol–water partitioning was used to conclude that it does not bind, but direct binding to membranes was not assessed²¹.

VSTX1 was isolated from tarantula venom based on its ability to inhibit the KvAP voltage-dependent K⁺ channel¹⁴. The binding site for toxin on KvAP must be confined to the voltage sensor because protein corresponding to the isolated voltage sensor is equally as effective as the entire KvAP channel in extracting VSTX1 and other voltage sensor toxins from the venom (V. Ruta and R. M.,

submitted). When 20 nM VSTX1 is applied to a 3.0 ml chamber corresponding to the external side of a planar lipid membrane formed on a 300 µm diameter hole, KvAP channels become inhibited to about 60% of their control value over a period of more than one hour (Fig. 1a). The experiment illustrates two key features of inhibition by voltage sensor toxins: high apparent affinity, in this case at equilibrium a half inhibitory concentration of about 30 nM, and very slow rates of equilibration.

VSTX1 binds to KvAP with a very reduced affinity when the channel is extracted from the membrane with detergent (Fig. 1b). Without the membrane the binding affinity is so low that our estimate of K_d is only approximate, between 100 µM and 500 µM. What accounts for the approximately 10⁴-fold discrepancy between toxin inhibition of channels in membranes and toxin binding to channels in the absence of the membrane? As shown below, this large discrepancy can in fact be fully accounted for by a favourable membrane–water partition coefficient for toxin, without invoking any other hypothetical causes such as an alteration of voltage sensor structure in detergent.

When large unilamellar phospholipid vesicles²² are mixed with a solution containing VSTX1 and then spun down by centrifugation, essentially all the toxin goes with the vesicles (Fig. 2a–c). Binding of VSTX1 to phospholipid vesicles does not require negatively charged lipids because the same result is observed whether negatively charged (POPE:POPG = 3:1, Fig. 2b) or neutral (POPC, Fig. 2c) phospholipids (see Methods) are used. The effect is unique to VSTX1 because in a control experiment we observe that a similar sized pore-blocking toxin from scorpion venom, agitoxin2 (AgTx2)

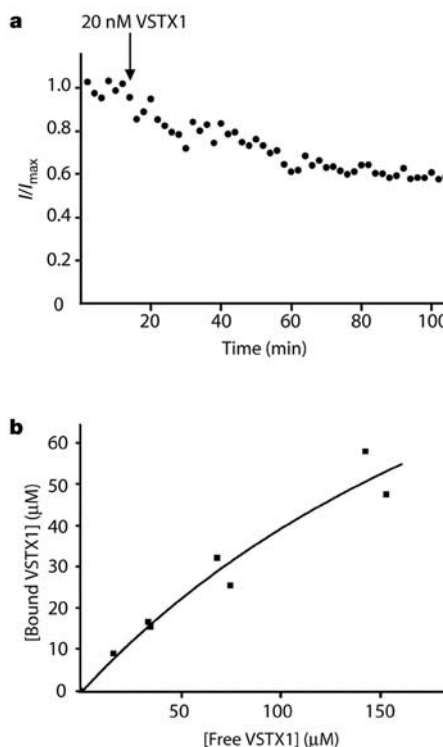


Figure 1 VSTX1 binds to the KvAP channel with high affinity in lipid membranes and low affinity in detergent micelles. **a**, Inhibition of KvAP by VSTX1 in membranes. Currents were elicited by depolarization to 100 mV for 200 ms every 120 s. Currents were normalized to the average control value. VSTX1 (20 nM) was added to the external solution at the point indicated by the arrow. **b**, Binding curve of VSTX1 to KvAP–Co²⁺ resin. The curve corresponds to equation: $[\text{Bound VSTX1}] = [\text{KvAP}] \times [1 + K_d/[\text{Free VSTX1}]]^{-1}$, with equilibrium dissociation constant $K_d = 303 \pm 208 \mu\text{M}$ and $[\text{KvAP}] = 159 \pm 79 \mu\text{M}$ (see Methods).

(ref. 23), shows no detectable binding to phospholipid vesicles (Fig. 2d, e).

To quantify the binding of VSTX1 to membranes we used intrinsic tryptophan fluorescence²⁴. VSTX1 contains three tryptophan residues. When small unilamellar phospholipid vesicles (POPE:POPG = 3:1) are added to VSTX1 the fluorescence spectrum undergoes a blue-shift and an increased maximal intensity characteristic of tryptophan transfer from an aqueous to a hydrophobic environment (Fig. 3a). As a control, Fig. 3b shows that zwitterionic tryptophan free in solution does not exhibit the spectral changes because by itself it does not partition into membranes; it has to be carried there by the protein²⁴. A graph of relative fluorescence intensity at 340 nm as a function of phospholipid concentration allows the quantification of VSTX1 partitioning into membranes. The partition coefficient is about 10^5 , corresponding to a transfer free energy of nearly 7 kcal mole⁻¹.

The approximately 10^4 -fold discrepancy between the K_d for VSTX1 binding to KvAP in the absence of membranes (100 μ M to 500 μ M) and the solution concentration at which VSTX1 inhibits KvAP in membranes (~ 10 nM) can be explained by the membrane-water partition coefficient. Direct protein-protein interactions between VSTX1 and KvAP indeed need not be strong because the toxin is accumulated at very high concentrations in the membrane. This explanation accounts for one of the enigmatic properties of voltage sensor toxins, their paradoxically small binding site on voltage-dependent K⁺ channels. A large fraction of the binding energy can be accounted for simply through the nonspecific free energy of membrane partitioning.

The structure of hanatoxin¹⁸ shows highlighted in green the hydrophobic face²¹ (Fig. 4a). A sequence alignment of several voltage sensor toxins shows that this face is fairly well conserved, as hydrophobic, in VSTX1 and other members of the toxin family (Fig. 4b). Two of three tryptophan residues on VSTX1 are clearly

located on the face, which presumably inserts into the hydrophobic core of the membrane to cause the blue shift in the fluorescence spectrum²⁴. Conservation of the hydrophobic face (as hydrophobic) suggests that other voltage sensor toxins may also partition into the membrane. The scanning mutagenesis study of SGTx1 is consistent with membrane partitioning: some mutations on the hydrophobic face may influence the effective inhibitory concentration by simply altering the free energy of membrane partitioning²⁰.

The membrane partitioning suggests that VSTX1 probably inhibits KvAP by first binding to the membrane and then diffusing laterally until it finds its binding site on the channel. A multi-step process of toxin insertion into the membrane, lateral diffusion and then binding to the voltage sensor complicates the interpretation of the slow kinetics of inhibition and recovery with voltage sensor toxins. When at equilibrium with an aqueous solution containing 10 nM VSTX1 the local toxin concentration in the membrane because of partitioning will be closer to 1.0 mM. This is an interesting realization because it suggests that the channel will be constantly bombarded at high frequency by toxin molecules in the membrane. The argument goes as follows. The membrane diffusion coefficient for a molecule such as the toxin is probably around 10^{-8} cm² s⁻¹ (ref. 25). Combining this with a local concentration near 1.0 mM, the collision frequency is expected to be orders of magnitude faster (occurring on the sub-millisecond timescale) than the observed kinetics of inhibition (occurring on the timescale of minutes, Fig. 1a) (ref. 26). The slow inhibition kinetics are as yet unexplained: they could be caused either by a slow rate of membrane partitioning, by poor efficiency of successful encounter in the toxin-channel binding step or by a combination of both effects. The demonstration of membrane partitioning represents an important step towards understanding the slow kinetics.

The observation that VSTX1 and possibly other voltage sensor toxins become concentrated in the membrane has important

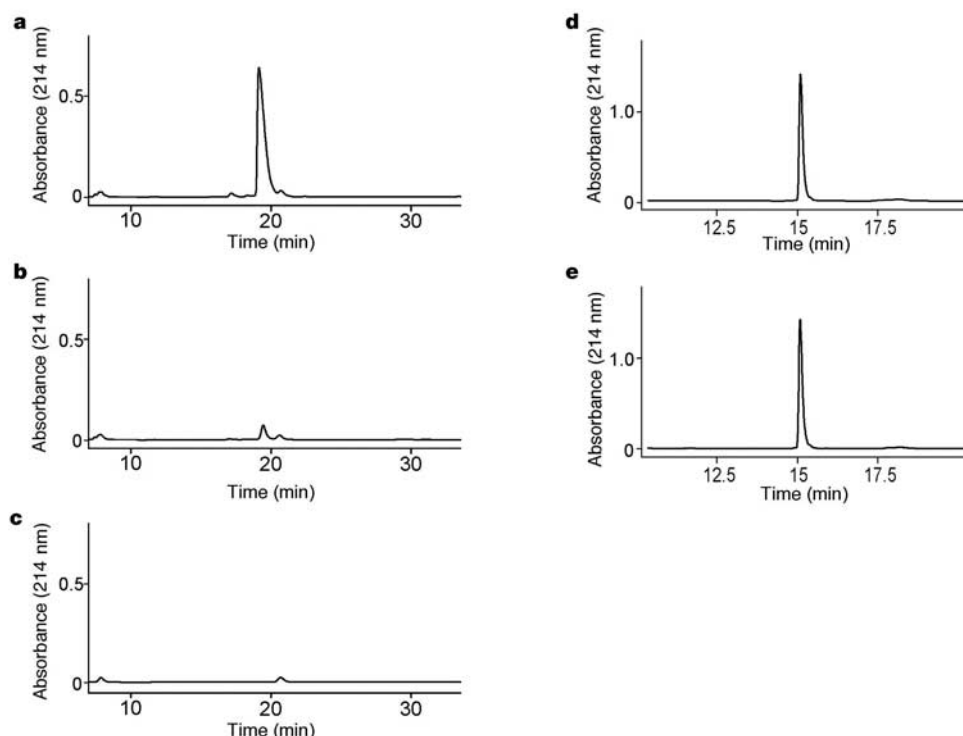


Figure 2 VSTX1 binds to lipid membranes. **a–c**, Reverse-phase HPLC profile of VSTX1 in the supernatant after ultra-centrifugation, in the absence of lipid membranes (**a**), after addition of POPE:POPG (3:1) vesicles (**b**), after addition of POPC vesicles (**c**).

d, e, Reverse-phase HPLC profile of supernatant after ultracentrifugation of control AgTx2 in the absence of lipid membranes (**d**) after mixing with POPE:POPG (3:1) vesicles (**e**) and ultra-centrifugation.

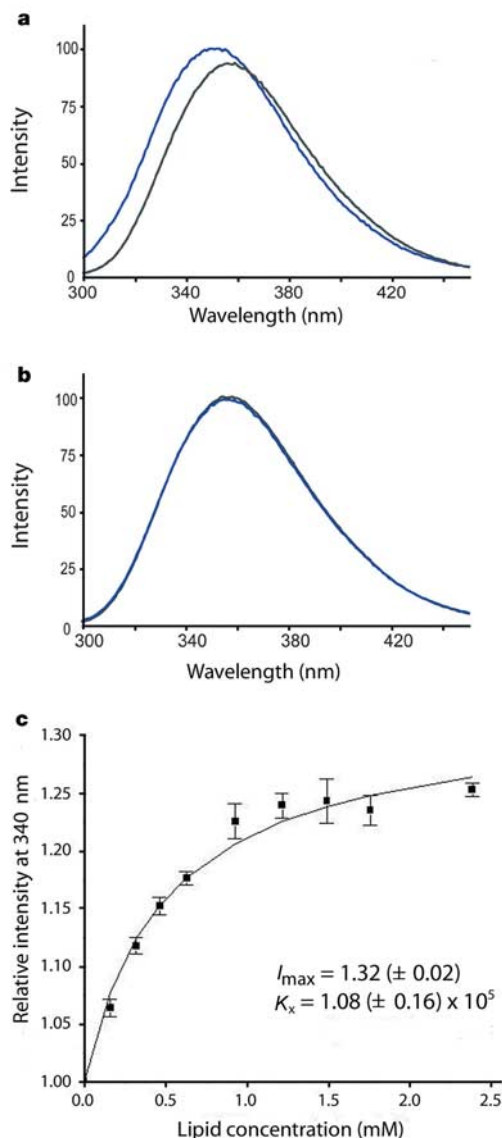


Figure 3 Increase in tryptophan fluorescence of VSTX1 on partitioning into membranes. **a**, Spectra of VSTX1 in the presence of (blue) and in the absence of lipid vesicles (black). **b**, Spectra of zwitterionic tryptophan fluorescence in the presence of (blue) and in the absence of lipid vesicles (black). The lipid concentration was 930 μ M for both **a** and **b**. **c**, Membrane-partitioning curve. Normalized fluorescence increase is plotted against average lipid concentration (see Methods). (Relative intensity at 340 nm, mean $\pm$ s.e.m., $n = 6$ for first four data points, $n = 3$ for the remaining data). Partition coefficient K_x is defined in Methods.

implications for the mechanism of toxin inhibition and also for the mechanism of gating in voltage dependent K^+ channels. It has been assumed that voltage sensor toxins remain bound to the voltage sensor as it moves between its closed and opened conformations²¹, but this is not necessarily so. High membrane toxin concentrations and the possibility of local (within the membrane) association and dissociation rates on the timescale of gating mean this assumption has to be re-evaluated. With regard to the mechanism of voltage-dependent gating, the clear implication of the present study is that the voltage sensor seems to be accessible from within the membrane but not from the aqueous solution. This finding is consistent with the conceptual model of voltage-dependent gating presented on the basis of KvAP studies^{5,6}.

The important conclusion from these experiments is that we see

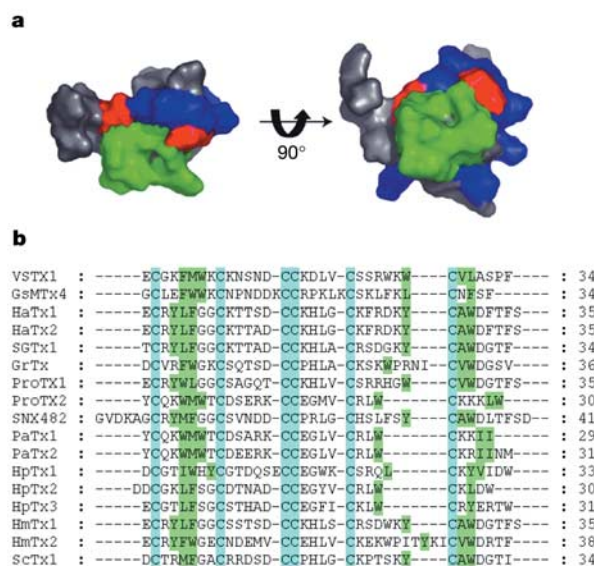


Figure 4 Hydrophobic face of voltage sensor toxins is conserved. **a**, Solution structure of hanatoxin (PDB ID:1D1H) in both front view (right) and side view (left). Hydrophobic face is coloured green, basic residues blue and acidic residues red. **b**, Sequence alignment of voltage sensor toxins. Residues that form the hydrophobic face are highlighted with green and cysteine residues are highlighted with cyan. Voltage sensor toxins are: VSTX1, voltage-sensor toxin 1; GsMTx4, *Grammostola* mechanotoxin number 4 (GI: 2541236); HaTx1 and HaTx2, hanatoxins (GI: 1245765 and GI: 1245766, respectively); SGTx1 (GI: 7388334); GrTx, grammotoxin (GI: 451235); ProTx1 and ProTx2 (GI: 25091452 and GI: 25091451, respectively); SNX482 (GI: 7388243); PaTx1 and PaTx2, phrixotoxins²⁹; HpTx1, HpTx2, and HpTx3, heteropodatoxins (GI: 17433177, GI: 17433178 and GI: 17433179, respectively); HmTx1 and HmTx2, heteroscolodratotoxins³⁰; ScTx1, stomatocin 1 (ref. 30).

how nature exploits membrane partitioning of a protein toxin to achieve high apparent affinity and to deliver the toxin to its voltage sensor target. It is well known that small organic compounds such as the poison arrow toxin batrachotoxin, alkaloid nerve toxins, pesticides such as DDT and some anaesthetic molecules alter voltage-dependent gating through the lipid membrane²⁷. It is fascinating now to understand that even a protein toxin whose function is to alter voltage-dependent gating acts in a similar manner. □

Methods

Electrophysiology

KvAP channels were reconstituted from decyl β -D-maltpyranoside (DM) into lipid vesicles as described for KcsA²⁸ and vesicle fusion into planar lipid bilayers was performed as described¹⁴. The recording chamber volume was 3.0 ml and the membrane diameter 300 μ m. In these assays the small membrane size ensures that toxin binding to the membrane does not deplete toxin concentration in solution. Membrane voltage was controlled and current recorded using an Axopatch 200B amplifier with a Digidata1322A analogue to digital converter and Axoclamp software (Axon Instruments).

Pull-down assay

VSTX1 was purified from *G. spatulata* spider venom and KvAP channel protein was expressed and purified as described¹⁴. Toxin binding was performed by first linking the KvAP channel to Co^{2+} resin through a hexahistidine tag on the carboxyl terminus. Co^{2+} resin (50 μ l) saturated with the KvAP channel was added to 200 μ l of buffer (20 mM Tris, pH 8.0, 5 mM DM and 100 mM KCl) containing a known concentration of toxin and mixed for 1 h. The resin was then spun down and the toxin quantity remaining in the supernatant was run on an Agilent 1100 Series high-performance liquid chromatography (HPLC) with a C-18 reverse-phase 5- μ m, 300- $\text{\AA}$ Vydac (218TP54) column and the area of peaks that corresponded to VSTX1 were integrated. For each measurement a control sample containing resin but without channel was used to determine nonspecific binding and a control without resin was used as a standard to correlate HPLC peak area and toxin concentration. To estimate the concentration of KvAP channels bound to Co^{2+} resin channels were eluted with 400 mM imidazole and the concentration measured at A_{280}

using extinction coefficient $33,543 \text{ M}^{-1} \text{ cm}^{-1}$. The K_d estimated by the above method was $303 \pm 208 \mu\text{M}$. We used a second method to estimate K_d by loading channel-saturated Co^{2+} resin on to a 100 μl column, applying toxin in a 100 μl volume and then washing the resin briefly with solution (above) containing 10 mM imidazole. The channel plus bound toxin was eluted in 400 mM imidazole and run on HPLC to determine bound toxin peak area. This second method allows direct determination of bound toxin and the smaller volumes allowed us to reach higher toxin concentrations. The column wash step probably allows some fraction of the toxin to dissociate. This method yielded a K_d of $917 \pm 81 \mu\text{M}$. Data using the first method are shown in Fig. 1b. Both methods report that toxin binds with very low affinity.

Spin-down assay

1-palmitoyl-2-oleoyl-phosphatidylethanolamine (POPE; 10 mg ml^{-1}) and 1-palmitoyl-2-oleoylphosphatidyl-DL-glycerol (POPG; 10 mg ml^{-1}) were mixed in a 3:1 ratio and were dried down under Argon gas. The lipid pellet was washed with pentane and suspended with a buffer (10 mM HEPES, pH 7.0 and 150 mM KCl) and sonicated. Large-unilamellar vesicles (LUV) were prepared by adding 5 mM DM (final concentration) to small-unilamellar vesicles (SUV) and incubating at room temperature for 2 h to induce fusion of SUVs²². Extensive dialysis was performed against 10 mM HEPES (pH 7.0) and 150 mM KCl to remove detergent. For 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphatidylcholine (POPC) vesicle preparation, 50 mM KCl was used instead of 150 mM KCl. Toxins (VSTX1 and AgTx2; $63 \mu\text{M}$) were mixed with approximately 9 mg ml^{-1} (total concentration) of LUVs and incubated for 30 min before ultra-centrifugation. Toxin-vesicle mixtures were centrifuged at 371,000 g for 80 min. LUVs were used for the spin-down assay to achieve a more complete removal of vesicles from solution. The resulting supernatants were run on the HPLC with a C-18 column. Solution A (H_2O , 0.1% trifluoroacetic acid, TFA) and solution B (90% acetonitrile, 10% H_2O , 0.1% TFA) were used in the HPLC run. VSTX1 was run on a 30–42% B gradient over 40 min and AgTx2²³ was run on a 0–73% B gradient over 30 min.

Fluorescence assay

Partitioning of VSTX1 into membranes was quantified by the fluorescence intensity increase of the tryptophan residues. Fluorescence measurements were collected on a Spex Fluorolog 3-11 spectrofluorometer with 270 nm excitation, 5 nm slit width (both emission and excitation), and cross-polarized configuration ($\text{Ex}_{\text{pol}} = 90^\circ$, $\text{Em}_{\text{pol}} = 0^\circ$) in a 3×3 -mm cuvette. Cross-polarized configuration was used to minimize vesicle scattering as described²⁴. The buffer contained 10 mM HEPES (pH 7.0) and 150 mM KCl. 20 mg ml^{-1} of POPE:POPG (3:1) vesicles in the buffer was prepared for the titration study. The concentration of VSTX1 and tryptophan were $3.3 \mu\text{M}$ and $10 \mu\text{M}$, respectively. The raw data were corrected by subtracting a scattering spectrum from a blank vesicle sample and by applying correction factors using tryptophan fluorescence as a reference²⁴. Fluorescence intensity was normalized to the value in the absence of vesicles. A mole-fraction partition coefficient (K_x) was calculated, based on the best fits of the data to the equation²⁴: $I([L]) = 1 + (I_{\text{max}} - 1)K_x[L]/([W] + K_x[L])$. I_{max} is the maximum fluorescence increase upon complete binding, $[L]$ is average lipid concentration (60% of total lipid concentration used for SUV), $[W]$ is the molar concentration of water (55.3 M).

Received 10 March; accepted 5 May 2004; doi:10.1038/nature02632.

- Miller, C., Moczydlowski, E., Latorre, R. & Phillips, M. Charybdotoxin, a protein inhibitor of single Ca^{2+} -activated K^+ channels from mammalian skeletal muscle. *Nature* **313**, 316–318 (1985).
- Hidalgo, P. & MacKinnon, R. Revealing the architecture of a K^+ channel pore through mutant cycles with a peptide inhibitor. *Science* **268**, 307–310 (1995).
- Imredy, J. P., Chen, C. & MacKinnon, R. A snake toxin inhibitor of inward rectifier potassium channel ROMK1. *Biochemistry* **37**, 14867–14874 (1998).
- Hugues, M., Romey, G., Duval, D., Vincent, J. P. & Lazdunski, M. Apamin as a selective blocker of the calcium-dependent potassium channel in neuroblastoma cells: voltage-clamp and biochemical characterization of the toxin receptor. *Proc. Natl Acad. Sci. USA* **79**, 1308–1312 (1982).
- Jiang, Y., Ruta, V., Chen, J., Lee, A. & MacKinnon, R. The principle of gating charge movement in a voltage-dependent K^+ channel. *Nature* **423**, 42–48 (2003).
- Jiang, Y. *et al.* X-ray structure of a voltage-dependent K^+ channel. *Nature* **423**, 33–41 (2003).
- Miller, C. Diffusion-controlled binding of a peptide neurotoxin to its K^+ channel receptor. *Biochemistry* **29**, 5320–5325 (1990).
- Escobar, L., Root, M. J. & MacKinnon, R. Influence of protein surface charge on the bimolecular kinetics of a potassium channel peptide inhibitor. *Biochemistry* **32**, 6982–6987 (1993).
- Goldstein, S. A., Pheasant, D. J. & Miller, C. The charybdotoxin receptor of a Shaker K^+ channel: peptide and channel residues mediating molecular recognition. *Neuron* **12**, 1377–1388 (1994).
- Ranganathan, R., Lewis, J. H. & MacKinnon, R. Spatial localization of the K^+ channel selectivity filter by mutant cycle-based structure analysis. *Neuron* **16**, 131–139 (1996).
- Swartz, K. J. & MacKinnon, R. Hanatoxin modifies the gating of a voltage-dependent K^+ channel through multiple binding sites. *Neuron* **18**, 665–673 (1997).
- Swartz, K. J. & MacKinnon, R. Mapping the receptor site for hanatoxin, a gating modifier of voltage-dependent K^+ channels. *Neuron* **18**, 675–682 (1997).
- Swartz, K. J. & MacKinnon, R. An inhibitor of the Kv2.1 potassium channel isolated from the venom of a Chilean tarantula. *Neuron* **15**, 941–949 (1995).
- Ruta, V., Jiang, Y., Lee, A., Chen, J. & MacKinnon, R. Functional analysis of an archaeobacterial voltage-dependent K^+ channel. *Nature* **422**, 180–185 (2003).
- Winterfield, J. R. & Swartz, K. J. A hot spot for the interaction of gating modifier toxins with voltage-dependent ion channels. *J. Gen. Physiol.* **116**, 637–644 (2000).
- Li-Smerin, Y. & Swartz, K. J. Gating modifier toxins reveal a conserved structural motif in voltage-gated Ca^{2+} and K^+ channels. *Proc. Natl Acad. Sci. USA* **95**, 8585–8589 (1998).
- Lee, C. W. *et al.* Solution Structure and Functional Characterization of SGTx1, a Modifier of Kv2.1 Channel Gating. *Biochemistry* **43**, 890–897 (2004).

- Takahashi, H. *et al.* Solution structure of hanatoxin1, a gating modifier of voltage-dependent K^+ channels: common surface features of gating modifier toxins. *J. Mol. Biol.* **297**, 771–780 (2000).
- Manoleras, N. & Norton, R. S. Three-dimensional structure in solution of neurotoxin III from the sea anemone *Anemonia sulcata*. *Biochemistry* **33**, 11051–11061 (1994).
- Wang, J. *et al.* Functional mapping of the molecular surface of a gating modifier toxin for voltage-gated K^+ channels. *Biophys. J.* **84**, 218a (2003).
- Lee, H. C., Wang, J. M. & Swartz, K. J. Interaction between extracellular Hanatoxin and the resting conformation of the voltage-sensor paddle in Kv channels. *Neuron* **40**, 527–536 (2003).
- Jiang, Q. X. *Spherical reconstruction: a novel method for structure determination from cryo-EM images of membrane proteins in small vesicles*. Thesis, Yale Univ. (2001).
- Garcia, M. L., Garcia-Calvo, M., Hidalgo, P., Lee, A. & MacKinnon, R. Purification and characterization of three inhibitors of voltage-dependent K^+ channels from *Leiurus quinquestriatus* var. *hebraeus* venom. *Biochemistry* **33**, 6834–6839 (1994).
- Ladokhin, A. S., Jayasinghe, S. & White, S. H. How to measure and analyze tryptophan fluorescence in membranes properly, and why bother? *Anal. Biochem.* **285**, 235–245 (2000).
- Almeida, P. F. F. & Vaz, W. L. C. In *Handbook of Biological Physics* (eds Lipowsky, R. & Sackmann, E.) 305–357 (Elsevier Science, Netherlands, 1995).
- Schurr, J. M. The role of diffusion in enzyme kinetics. *Biophys. J.* **10**, 717–727 (1970).
- Hille, B. *Ion Channels of Excitable Membranes* 635–662 (Sinauer Associates, Sunderland, MA, 2001).
- Heginbotham, L., LeMasurier, M., Kolmakova-Partensky, L. & Miller, C. Single streptomyces lividans K^+ channels: functional asymmetries and sidedness of proton activation. *J. Gen. Physiol.* **114**, 551–560 (1999).
- Diochot, S., Drici, M. D., Moinier, D., Fink, M. & Lazdunski, M. Effects of phrixotoxins on the Kv4 family of potassium channels and implications for the role of Ito1 in cardiac electrogenesis. *Br. J. Pharmacol.* **126**, 251–263 (1999).
- Escoubas, P., Diochot, S., Celerier, M. L., Nakajima, T. & Lazdunski, M. Novel tarantula toxins for subtypes of voltage-dependent potassium channels in the Kv2 and Kv4 subfamilies. *Mol. Pharmacol.* **62**, 48–57 (2002).

Acknowledgements We thank D. King for toxin synthesis for initial experiments, F. Valiyaveetil for help with fluorescence experiments, V. Ruta for help with toxin purification and electrophysiology and S. Long for providing AgTx2. This work was supported by a National Institutes of Health grant to R.M. R.M. is an investigator in the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.M. (mackinn@rockefeller.edu).

Bilayer-dependent inhibition of mechanosensitive channels by neuroactive peptide enantiomers

Thomas M. Suchyna¹, Sonya E. Tape², Roger E. Koeppe II³,
Olaf S. Andersen², Frederick Sachs¹ & Philip A. Gottlieb¹

¹Department of Physiology and Biophysics, SUNY at Buffalo, Buffalo, New York 14214, USA

²Department of Physiology and Biophysics, Weill Medical College of Cornell University, New York, New York 10021, USA

³Department of Chemistry and Biochemistry, University of Arkansas, Fayetteville, Arkansas 72701, USA

The peptide GsMTx4, isolated from the venom of the tarantula *Grammostola spatulata*, is a selective inhibitor of stretch-activated cation channels (SACs)¹. The mechanism of inhibition remains unknown; but both GsMTx4 and its enantiomer, enGsMTx4, modify the gating of SACs, thus violating a trademark of the traditional lock-and-key model of ligand–protein interactions. Suspecting a bilayer-dependent mechanism, we examined the effect of GsMTx4 and enGsMTx4 on gramicidin A (gA) channel gating². Both peptides are active, and the effect increases with the degree of hydrophobic mismatch between bilayer thickness and channel length, meaning that GsMTx4 decreases the energy required to deform the boundary lipids adjacent to the channel. GsMTx4 decreases inward SAC single-channel currents but has no effect on outward currents, suggesting it is located within a Debye length of the outer

vestibule of the SACs, but significantly farther from the inner vestibule. Likewise, GsMTx4 decreases gA single-channel currents. Our results suggest that modulation of membrane proteins by amphipathic peptides—mechanopharmacology—involves not only the protein itself but also the surrounding lipids. The surprising efficacy of the D form of GsMTx4 peptide has important therapeutic implications, because D peptides are not hydrolysed by endogenous proteases and may be administered orally.

GsMTx4 selectively inhibits the gating of cation selective SACs that derive their gating energy from bilayer tension. The gating of these SACs is known to be modified by amphiphiles such as arachidonic acid, general anaesthetics^{3,4} and lysophospholipids⁵. These amphiphiles alter bilayer mechanical properties as assayed by their effects on gA channels^{2,6,7}, which themselves are sensitive to bilayer tension⁸. GsMTx4 itself is amphipathic⁹: the NMR structure⁹ showed an inhibitory cysteine-knot motif with a net charge of +5 (the carboxyl terminus is amidated)¹⁰, and a hydrophobic face

(Fig. 1a). If GsMTx4 adsorbs to lipid bilayers, it may perturb the lipid packing adjacent to a channel (Fig. 1b) and thereby alter the line tension of the SAC/bilayer interface¹¹—and thus SAC gating without binding to the channel. If so, a similar mechanism would apply to gA channels (Fig. 1c).

GsMTx4 inhibits ubiquitous cationic SACs in a variety of vertebrate cell types (200–600 nM, for example, chick heart, rat astrocytes and skeletal muscle, and human smooth muscle), despite the fact that they have different kinetic properties and presumably different structure. It inhibits channel activity when applied to the extracellular face of the cell membrane by increasing the membrane tension required for activation (Fig. 2a, b), suggesting that GsMTx4 acts as a gating modifier. In the presence of GsMTx4 (Fig. 2c–e), the inward currents were reduced ~10% with no observable effect on outward currents. This effect is predicted by Poisson–Boltzmann models (T.M.S., P.A.G. and F.S., unpublished observations), when GsMTx4 is located within 1.5 nm of the channel vestibule.

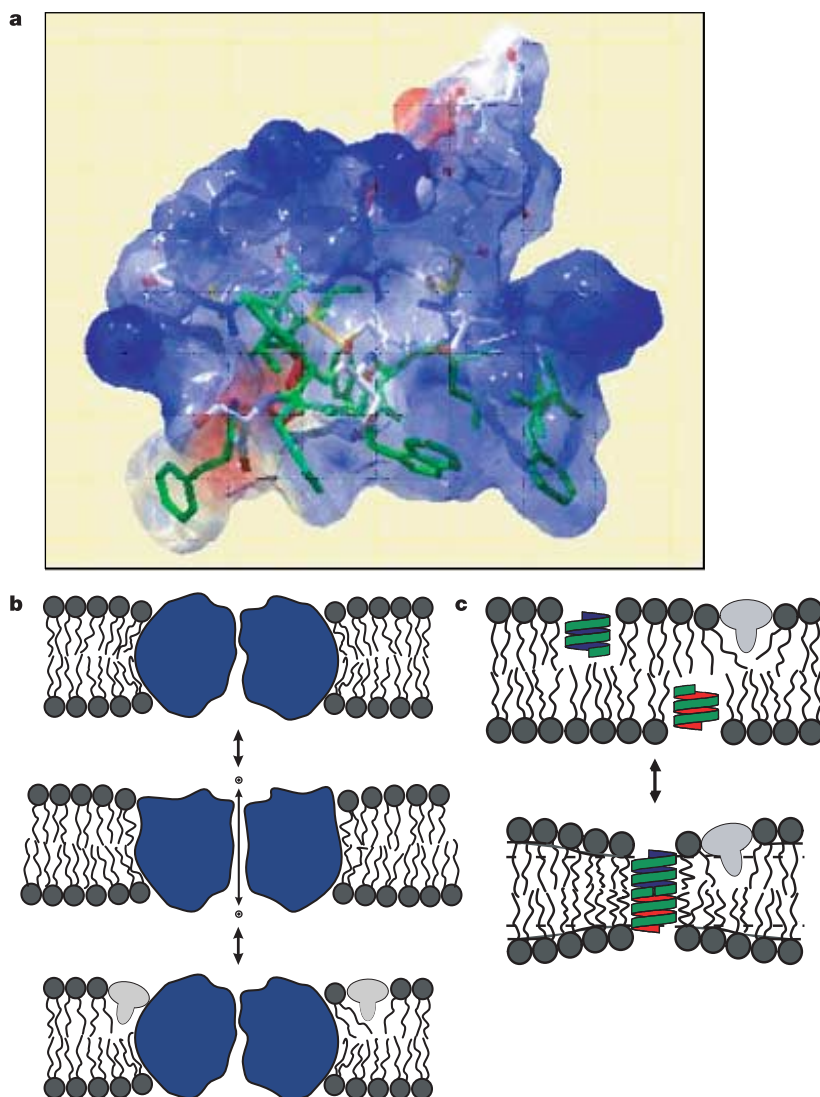


Figure 1 GsMTx4 and its possible bilayer-modifying effects. **a**, Rendered image of GsMTx4 coloured by Coulomb electrostatic potential at $2.8 k_B T/e$ (k_B is the Boltzmann constant, T temperature in K and e the elementary charge). Positive residues (blue); negative residues (red); hydrophobic residues (green). **b**, Schematic diagram of how GsMTx4 may inhibit SACs by perturbing the channel/bilayer boundary without necessarily

being in physical contact with the channel. **c**, Schematic diagram of gA channel formation, which involves a local bilayer thinning. The bilayer deformation energy could be reduced if GsMTx4 were near/at the channel, which would increase the dimerization constant (channel activity).

Although GsMTx4, unlike amphiphiles such as lysophospholipids, does not alter patch capacitance¹², we nevertheless suspected that GsMTx4 could function by altering bilayer properties. Using the tryptophan fluorescence¹³ of GsMTx4, we measured its binding to large unilamellar lipid vesicles. The partition coefficient, defined as $([\text{GsMTx4}]_{\text{bil}}/[\text{Lipid}])/([\text{GsMTx4}]_{\text{aq}}/[\text{H}_2\text{O}])$ (where the subscripts denote the bilayer and aqueous phases, respectively, for example see ref. 14), was $\sim 3 \times 10^5$ —corresponding to an adsorption coefficient of 10^{-3} cm (defined as $\{\text{GsMTx4}\}_{\text{bil}}/[\text{GsMTx4}]_{\text{aq}}$, where $\{\text{GsMTx4}\}_{\text{bil}}$ denotes the surface density, in moles per unit area). GsMTx4 thus binds avidly to lipid bilayers: at 500 nM, (the

effective K_D for gating modification) the GsMTx4:lipid ratio in the outer leaflet is $\sim 1:300$, meaning that the peptide concentration within a 1 nm thick layer of membrane at the bilayer/solution interface is 10^4 -fold higher than the bulk concentrations. This has two implications. First, the intrinsic affinity between a lipid-bound GsMTx4 and the channel could be quite modest and yet account for the observed K_D ¹⁵. Second, the mean surface density of GsMTx4 is sufficiently high that fast exchange ($\sim 50 \mu\text{s}$) between the bulk bilayer and the lipids adjacent to the channel can alter the local (and global) bilayer mechanical properties. The apparent association and dissociation rates for SAC effects ($k_a \approx 10^{-6} \text{ M}^{-1} \text{ s}^{-1}$,

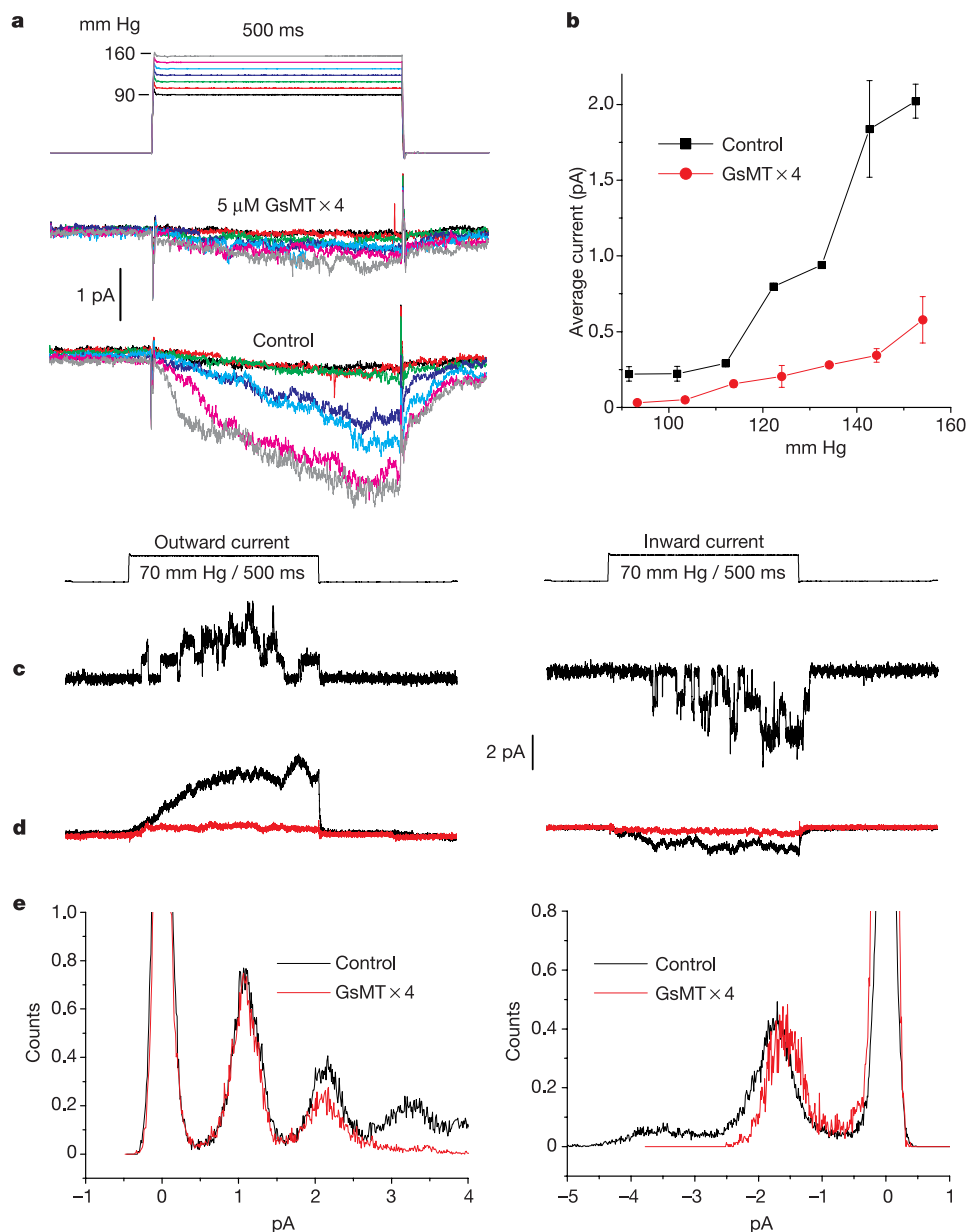


Figure 2 GsMTx4 increases the pipette pressure (membrane tension) required for activation and is adjacent to the channel. **a**, Ensemble-averaged SAC currents from a single outside-out astrocyte patch in the absence and presence of 5 μM GsMTx4. **b**, Average patch currents as a function of pressure in the absence and presence of 5 μM GsMTx4 (2 patches, mean $\pm$ s.e.m.) shows that pressure can overcome peptide inhibition. The stimulus is near saturation at the highest pressure attainable without lysis. GsMTx4 suppressed the average inward current by $\sim 80\%$, but the open state dwell time distribution was not affected (data not shown). **c**, Single-channel currents from a chick

heart cell, the outside-out patch shows mild rectification in the absence of GsMTx4. **d**, GsMTx4 reduces average currents in the chick heart cell outside-out patch. Control (black trace) and 5 μM GsMTx4 (red trace) show inhibition, that is slightly voltage dependent, reducing inward currents by $\sim 70\%$ and outward currents by $>90\%$. **e**, Histograms of unitary current amplitude showing no change in outward current amplitude (left) (+50 mV pipette potential: control, 1.10 pA; 5 μM GsMTx4, 1.08 pA), but a 10% reduction in the inward current amplitude (right) (-50 mV pipette potential: control, 1.74 pA; 5 μM GsMTx4, 1.58 pA).

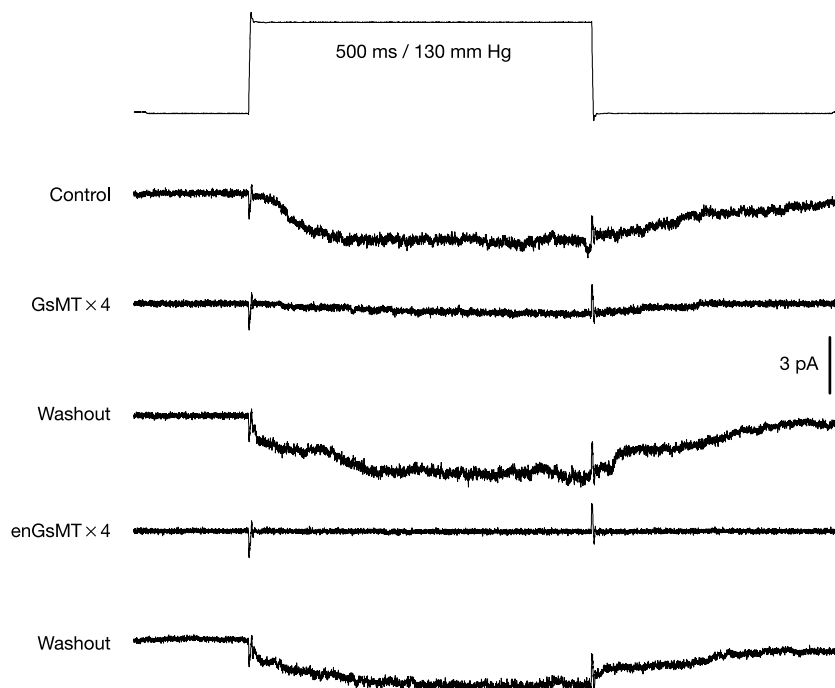


Figure 3 GsMTx4 and enGsMTx4 both inhibit SACs. The averaged current from a rat astrocyte outside-out patch is inhibited by both GsMTx4 and enGsMTx4 (5 μ M).

$k_d \approx 0.2 \text{ s}^{-1}$)¹ are consistent with such a fast exchange (as well as slower, more specific binding).

To examine whether GsMTx4 could alter SAC function through a purely bilayer-dependent mechanism, we synthesized enGsMTx4 from D amino acids, and tested its effect on SACs. Except for its optical properties, enGsMTx4 is indistinguishable from wild-type GsMTx4. Surprisingly, enGsMTx4 inhibits SACs with similar efficacy to GsMTx4 (Fig. 3). Thus, GsMTx4 is unlikely to alter SAC gating by a lock-and-key mechanism, but, judging by the effects on conductance (Fig. 2e), the peptide is in close proximity to the channel.

To examine the effects of GsMTx4 on the constitutive properties of lipid bilayers, we examined how it modifies the gating of gA channels in planar lipid bilayers, as gA has proven useful as a probe of bilayer mechanics^{2,8,16}. gA channels are dimers formed by the transbilayer association of monomers in opposing monolayers¹⁷. Channel formation is generally associated with local bilayer thinning (Fig. 1c)^{18,19}, which makes gA monomer $\leftrightarrow$ dimer equilibrium and thus the channel lifetime and appearance rate sensitive to bilayer thickness, elastic moduli and curvature stress. Thus, if GsMTx4 binds to the lipid bilayer, one would expect it to also alter the lipid packing adjacent to the channel and thus the gA gating kinetics and, because it is cationic decrease the channel current.

Figure 4 shows that GsMTx4 increases the channel activity of gA (and gA analogues of different length) (Fig. 4a), and reduces the single-channel current (Fig. 4b). The increased activity results from a 10–25-fold increase in appearance rate, and an approximately 2-fold increase in open channel lifetime (Fig. 4c), with channels formed by the shorter gA analogues being more sensitive than those formed by the longer analogues. These changes in channel function do not result from specific binding of GsMTx4 to gA. First, the changes in channel lifetime do not depend on the channels' helix sense, as shown using a pair of enantiomeric gA analogues (Fig. 4d)²⁰. Second, GsMTx4 most likely resides at the bilayer/

solution interface, meaning that any channel–peptide interactions would be expected to involve the four tryptophan residues at the channel/solution interfaces; but channels formed by a tryptophan $\rightarrow$ tyrosine substituted gA analogue²¹ are affected similarly as other gA channels (Fig. 4d). Third, GsMTx4 and enGsMTx4 similarly modulate gA activity, as 400 nM enGsMTx4 increases the channel appearance rates >10 -fold and the lifetime 4-fold. If there is any difference between GsMTx4 and enGsMTx4, it is that enGsMTx4 is more potent, as also may be the case with SACs (Fig. 3). We conclude that GsMTx4 exerts its effects on gA channels (at the same concentrations that affect SACs)¹ by altering the lipid packing at the bilayer/solution/channel interface.

The similarity of the GsMTx4 and enGsMTx4 effects is difficult to reconcile with traditional mechanisms for peptide–protein interactions, which involve stoichiometric, sterically well-defined, binding reactions. The changes in gA channel function are readily explained by altered bilayer properties, which alter the organization in the local environment of the channel and thus the free energy difference for the gA monomer $\leftrightarrow$ dimer equilibrium^{6,7,22}. Although the conformational changes underlying SAC gating are different from those underlying gA channel gating, the similar effects of GsMTx4 and enGsMTx4 on SACs, and the ability of GsMTx4 to modify SACs from different tissues, suggests that the mode of action is the same—altered local bilayer mechanics. The efficacy of the D form of GsMTx4 is important for clinical applications. For example, GsMTx4 can inhibit cardiac arrhythmias²³ and D peptides are resistant to proteolysis, which may permit oral administration.

Our results have implications for interpreting channel and other membrane protein function, as similar mechanopharmacological mechanisms undoubtedly apply to other mechanosensitive membrane proteins such as phospholipase A²⁴, G proteins²⁵ and voltage-dependent channels^{22,26}. Perhaps we should view 'mechanosensitive' membrane proteins also as receptors for amphipathic second messengers. Given the apparent generality of the mechanism of

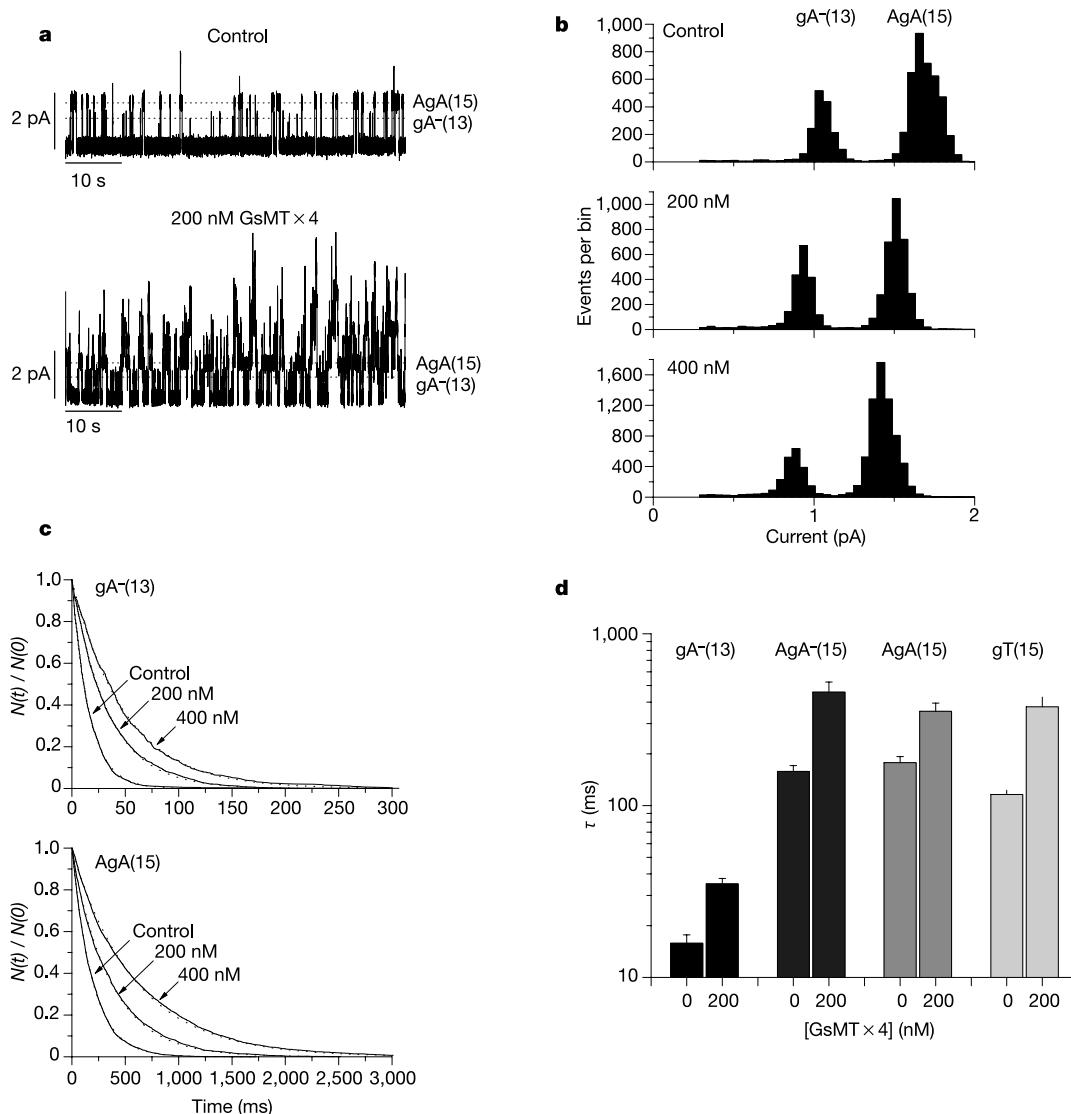


Figure 4 Effect of GsMTx4 on gA channels. **a**, Current traces before and after addition of 200 nM GsMTx4 to both sides of a bilayer doped with chain-shortened des-Val¹-Gly²-gA (gA⁻⁽¹³⁾) and [Ala¹]gA (AgA(15)) (numbers in parentheses denote number of residues in the sequence); gA⁻ is the enantiomer of gA⁺. **b**, **c**, Single-channel currents are reduced (**b**) and lifetimes increased (**c**) as a function of GsMTx4 concentration (see

Methods). **d**, Average lifetime (τ) as a function of channel length, helix sense and identity of the aromatic residues at the channel/solution interface. gA⁻⁽¹³⁾ and AgA(15) as above, AgA⁻⁽¹⁵⁾ is [D-Ala¹]gA⁻, gT is [Tyr^{9,11,13,15}]gA (ref. 21). 200 nM GsMTx4; columns and error bars denote mean $\pm$ s.d. ($n \geq 3$); 100 mV applied potential.

GsMTx4 action, it is surprising that GsMTx4 is highly selective for SACs¹. Most likely the distinguishing property of some SACs—their sensitivity to the bilayer tension—also makes them sensitive to the local lipid environment. The conceptual breakdown of the lock-and-key model does not seem to rule out specificity, but this break in tradition causes us to re-evaluate the notion of a binding site. □

Methods

Gramicidin experiments

gA analogues of different lengths and helix sense were synthesized and purified as described by Greathouse *et al.*²⁷. Single-channel measurements were performed at $25 \pm 1^\circ\text{C}$ in diphytanoylphosphatidylcholine (Avanti Polar Lipids)/*n*-decane bilayers, separating 1.0 M NaCl solutions buffered to pH 7 using 0.01 M HEPES using the bilayer punch²⁸ and an AxoPatch 1B patch clamp. The gA analogues and GsMTx4 were added to both sides of the bilayer; the amount of gA added was adjusted such that the channel appearance rates before peptide addition were 0.1 to 1 s⁻¹. Single-channel current transitions were detected using a transition-based algorithm²⁸, from which single-channel current amplitude histograms and channel duration histograms were constructed.

Survivor plots of the channel lifetimes were fitted by a single exponential distribution $N(t)/N(0) = \exp(-t/\tau)$, where $N(t)$ is the number of channels with duration longer than time t , and τ is the average lifetime. All results are reported as mean $\pm$ standard deviation based on at least three experiments.

Synthesis and folding of enGsMTx4

The linear sequence of GsMTx4 having D amino acids was synthesized by SynPep. Purification and folding was performed as described for the L-form¹⁰. The crude material (125 mg) was dissolved in water and insoluble material was removed by centrifugation. The peptide was purified by RP-HPLC (reverse phase-high-performance liquid chromatography) (Zorbax 300 SB-C18, 9.4 $\times$ 250 mm column) with a 20% to 45% acetonitrile–water gradient, containing 0.1% trifluoroacetic acid, at a flow rate of 3 ml min⁻¹. Elution was monitored at 220 and 280 nm; the desired peptide fractions were identified by MALDI-TOF (matrix-assisted laser desorption/ionisation–time of flight), pooled and evaporated to dryness. The purified peptide (200 μ g) was dissolved in 800 μ l water plus 100 μ l 1 M Tris at pH 8.0. For folding, 50 μ l of reduced glutathione (5 mM; Sigma), and 50 μ l of oxidized glutathione (0.5 mM; Sigma) were added, and incubated for 24 h at room temperature. The reaction was terminated with ~ 1 μ l concentrated phosphoric acid. The folded peptide was purified with the gradient and detection described above using a C18 column (5 μ m, 40 $\times$ 250 mm) at a flow rate of 1 ml min⁻¹. The MALDI-TOF and retention time on analytical HPLC for enGsMTx4 was identical to

GsMTx4. The circular dichroism spectrum of 10 μ M enGsMTx4 in 5 mM phosphate buffer at pH 7.4 was inverted relative to GsMTx4 (ref. 10).

SAC analysis in rat astrocytes and chick heart cells

Primary reactive rat astrocytes cultures (passages 4 to 10) were prepared as previously described¹. Primary cultures of chick heart cells were prepared from 11-day-old embryos as previously described²⁹.

Patch clamp

Outside-out patches were formed using pipettes containing 140 mM KCl, 2 mM MgCl₂, 10 mM HEPES and 0.5 mM EGTA (as K⁺ salts). SAC currents were recorded using an Axopatch 200B patch clamp, with pClamp9 software and a Digidata 1322A acquisition system (Axon Instruments). Pressure was applied to the patch pipette with an HSPC-1 pressure clamp, and the patch was perfused using an ALA MP285 perfusion system (ALA Scientific Instruments). GsMTx4 was dissolved in the perfusion solution: 140 mM NaCl, 10 mM Na-HEPES, 5 mM KCl, and 1 mM MgCl₂. All experiments were repeated >2–3 times.

Partition coefficient

Fluorescence measurements were performed as described by Ladokhin *et al.*¹³ with an AMINCO-Bowman Series 2 spectrofluorimeter. The excitation and emission wavelengths were 280 nm (8 nm band pass, 0° polarizer) and 346 nm (16 nm band pass, 90° polarizer). Large unilamellar vesicles of egg lecithin (Avanti Polar Lipids) were formed by extrusion. Tryptophan (Sigma) and GsMTx4 titration experiments were performed in 8 mM phosphate buffer containing 1 mM EGTA at pH 7. Three spectra were collected at each lipid concentration with both tryptophan and GsMTx4 (both at 30 μ M). After background subtraction, the mean fluorescence intensity at each lipid concentration was corrected for vesicle scattering¹³ and normalized to the zero lipid fluorescence intensity. The results are reported assuming a vesicle diameter of 130 nm, and area per lipid in the monolayer of 0.7 nm².

Received 9 April; accepted 3 June 2004; doi:10.1038/nature02743.

- Suchyna, T. M. *et al.* Identification of a peptide toxin from *Grammostola spatulata* spider venom that blocks cation-selective stretch-activated channels. *J. Gen. Physiol.* **115**, 583–598 (2000).
- Andersen, O. S. *et al.* Ion channels as tools to monitor lipid bilayer-membrane protein interactions: gramicidin channels as molecular force transducers. *Methods Enzymol.* **294**, 208–224 (1999).
- Patel, A. J. *et al.* Inhalational anesthetics activate two-pore-domain background K⁺ channels. *Nature Neurosci.* **2**, 422–426 (1999).
- Patel, A. J., Lazdunski, M. & Honore, E. Lipid and mechano-gated 2P domain K⁺ channels. *Curr. Opin. Cell Biol.* **13**, 422–427 (2001).
- Perozo, E., Kloda, A., Cortes, D. M. & Martinac, B. Physical principles underlying the transduction of bilayer deformation forces during mechanosensitive channel gating. *Nature Struct. Biol.* **9**, 696–703 (2002).
- Lundbak, J. A. & Andersen, O. S. Lysophospholipids modulate channel function by altering the mechanical properties of lipid bilayers. *J. Gen. Physiol.* **104**, 645–673 (1994).
- Hwang, T. C., Koeppe, R. E. II & Andersen, O. S. Genistein can modulate channel function by a phosphorylation-independent mechanism: importance of hydrophobic mismatch and bilayer mechanics. *Biochemistry* **42**, 13646–13658 (2003).
- Goulian, M. *et al.* Gramicidin channel kinetics under tension. *Biophys. J.* **74**, 328–337 (1998).
- Oswald, R. E., Suchyna, T. M., McFeeters, R., Gottlieb, P. & Sachs, F. Solution structure of peptide

- toxins that block mechanosensitive ion channels. *J. Biol. Chem.* **277**, 34443–34450 (2002).
- Ostrow, K. L. *et al.* cDNA sequence and *in vitro* folding of GsMTx4, a specific peptide inhibitor of mechanosensitive channels. *Toxicon* **42**, 263–274 (2003).
- Markin, V. S. & Sachs, F. Thermodynamics of mechanosensitivity. *Physical Biol.* (in the press).
- Suchyna, T. & Sachs, F. Dynamic regulation of mechanosensitive channels: capacitance used to monitor patch tension in real time. *Phys. Biol.* **1**, 1–18 (2004).
- Ladokhin, A. S., Jayasinghe, S. & White, S. H. How to measure and analyze tryptophan fluorescence in membranes properly, and why bother? *Anal. Biochem.* **285**, 235–245 (2000).
- White, S. H., Wimley, W. C., Ladokhin, A. S. & Hristova, K. Protein folding in membranes: determining energetics of peptide-bilayer interactions. *Methods Enzymol.* **295**, 62–87 (1998).
- Kim, J., Mosior, M., Chung, L. A., Wu, H. & McLaughlin, S. Binding of peptides with basic residues to membranes containing acidic phospholipids. *Biophys. J.* **60**, 135–148 (1991).
- Lundbak, J. A. & Andersen, O. S. Spring constants for channel-induced lipid bilayer deformations—estimates using gramicidin channels. *Biophys. J.* **76**, 889–895 (1999).
- O’Connell, A. M., Koeppe, R. E. II & Andersen, O. S. Kinetics of gramicidin channel formation in lipid bilayers: transmembrane monomer association. *Science* **250**, 1256–1259 (1990).
- Elliott, J. R., Needham, D., Dilger, J. P. & Haydon, D. A. The effects of bilayer thickness and tension on gramicidin single-channel lifetime. *Biochim. Biophys. Acta* **735**, 95–103 (1983).
- Huang, H. W. Deformation free energy of bilayer membrane and its effect on gramicidin channel lifetime. *Biophys. J.* **50**, 1061–1070 (1986).
- Koeppe, R. E. II *et al.* On the helix sense of gramicidin A single channels. *Proteins* **12**, 49–62 (1992).
- Trudelle, Y. & Heitz, F. Synthesis and characterization of Tyr(Bzl)^{9,11,13,15} and Tyr^{9,11,13,15} gramicidin A. *Int. J. Pept. Protein Res.* **30**, 163–169 (1987).
- Lundbak, J. A. *et al.* Regulation of sodium channel function by bilayer elasticity the importance of hydrophobic coupling: effects of micelle-forming amphiphiles and cholesterol. *J. Gen. Physiol.* **123**, 599–621 (2004).
- Bode, F., Sachs, F. & Franz, M. R. Tarantula peptide inhibits atrial fibrillation. *Nature* **409**, 35–36 (2001).
- Lehtonen, J. Y. & Kinnunen, P. K. Phospholipase A2 as a mechanosensor. *Biophys. J.* **68**, 1888–1894 (1995).
- Gudi, S., Nolan, J. P. & Frangos, J. A. Modulation of GTPase activity of G proteins by fluid shear stress and phospholipid composition. *Proc. Natl Acad. Sci. USA* **95**, 2515–2519 (1998).
- Laitko, U. & Morris, C. E. Membrane tension accelerates rate-limiting voltage-dependent activation and slow inactivation steps in a *Shaker* channel. *J. Gen. Physiol.* **123**, 135–154 (2004).
- Greathouse, D. V., Koeppe, R. E. II, Providence, L. L., Shobana, S. & Andersen, O. S. Design and characterization of gramicidin channels. *Methods Enzymol.* **294**, 525–550 (1999).
- Andersen, O. S. Ion movement through gramicidin A channels. Single-channel measurements at very high potentials. *Biophys. J.* **41**, 119–133 (1983).
- Bett, G. C. & Sachs, F. Activation and inactivation of mechanosensitive currents in the chick heart. *J. Membr. Biol.* **173**, 237–254 (2000).

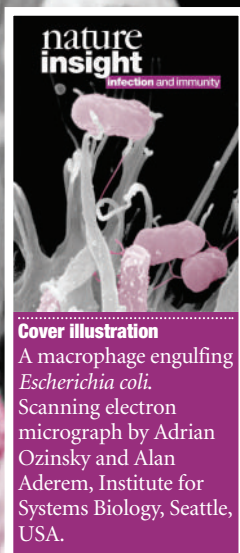
Acknowledgements We would like to thank J. Niggel (SUNY) and A. Ladokhin (UC Irvine) for the fluorescence data, and M. Teeling for the tissue culture. This work is supported by NIH grants to O.S.A. and F.S.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to O.S.A. (sparre@med.cornell.edu) or P.A.G. (philgott@buffalo.edu).

nature insight

infection and immunity



Cover illustration

A macrophage engulfing *Escherichia coli*. Scanning electron micrograph by Adrian Ozinsky and Alan Aderem, Institute for Systems Biology, Seattle, USA.

Infectious diseases remain a serious medical burden in both developing and industrialized countries. The emergence of new diseases such as HIV/AIDS and severe acute respiratory syndrome (SARS), the resurgence of known diseases such as West Nile disease and multidrug-resistant tuberculosis, and the threat of deliberately developed man-made infections are cause for concern. When vaccines or effective treatments are not available, we rely on the immune system to clear the host of infectious agents and disease. A better understanding of the tactics used by pathogens and of the immune system's defensive armoury is needed to pave the way for improved strategies of prevention and therapy.

The four articles in this Insight provide a snapshot of the issue, addressing in turn the nature and spread of emerging and re-emerging infectious diseases, the assault strategies used by major pathogens, the immediate innate immune response, and the antigen-specific acquired (or adaptive) immune response, which is the host's second line of defence.

David Morens, Gregory Folkers and Anthony Fauci review new and old infectious diseases that pose challenges for public health in the twenty-first century. Scott Merrell and Stanley Falkow discuss the strategies that pathogens use to evade the host's defences and secure their own survival. Bruce Beutler reviews the role of Toll-like receptors (TLRs) in early detection of and defence against pathogens in relation to clinical applications, and discusses how TLRs may contribute to autoimmunity and sterile inflammation. And Marianne Boes and Hidde Ploegh review successive stages of T-cell activation from a cell-biological perspective, discussing how functional data obtained *in vitro* may translate to an *in vivo* setting.

These articles provide an overview of the current status of each field, and consider directions for future research. We are indebted to the authors who contributed to this Insight, and we hope that you will find the reviews informative and stimulating.

Ursula Weiss Senior Editor

review articles:

242 The challenge of emerging and re-emerging infectious diseases

D. M. Morens,
G. K. Folkers
& A. S. Fauci

250 Frontal and stealth attack strategies in microbial pathogenesis

D. S. Merrell & S. Falkow

257 Inferences, questions and possibilities in Toll-like receptor signalling

B. Beutler

264 Translating cell biology *in vitro* to immunity *in vivo*

M. Boes & H. L. Ploegh

Editor, *Nature*: Philip Campbell
Publisher: Sarah Greaves
Insights Editor: Lesley Anson
Editorial Assistant: Timothy Gibbs
Production Editor: Rebecca Craven

Art Editor: Martin Harrison
Layouts: Alisdair Macdonald
Production: Sue Gray
Marketing: Claire Aspinall
Sponsorship: Claire Hines

Diagrams: Charlotte Cartwright
Radha Clelland
Nik Spencer
Mick Ayres

The challenge of emerging and re-emerging infectious diseases

David M. Morens, Gregory K. Folkers & Anthony S. Fauci

National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland 20892-2520, USA (e-mail: afauci@niaid.nih.gov)

Infectious diseases have for centuries ranked with wars and famine as major challenges to human progress and survival. They remain among the leading causes of death and disability worldwide. Against a constant background of established infections, epidemics of new and old infectious diseases periodically emerge, greatly magnifying the global burden of infections. Studies of these emerging infections reveal the evolutionary properties of pathogenic microorganisms and the dynamic relationships between microorganisms, their hosts and the environment.

Emerging infections (EIs) can be defined as “infections that have newly appeared in a population or have existed previously but are rapidly increasing in incidence or geographic range”¹. EIs have shaped the course of human history and have caused incalculable misery and death. In 1981, a new disease — acquired immune deficiency syndrome (AIDS) — was first recognized. As a global killer, AIDS now threatens to surpass the Black Death of the fourteenth century and the 1918–1920 influenza pandemic, each of which killed at least 50 million people^{2,3}. Of the ‘newly emerging’ and ‘re-emerging/resurging’ diseases that have followed the appearance of AIDS (Fig. 1), some have been minor curiosities, such as the 2003 cases of monkeypox imported into the United States⁴, whereas others, such as severe acute respiratory syndrome (SARS), which emerged in the same year⁵, have had a worldwide impact. The 2001 anthrax bioterrorist attack in the United States⁶ falls into a third category: ‘deliberately emerging’ diseases. EIs can be expected to remain a considerable challenge for the foreseeable future. Here we examine the nature and scope of emerging and re-emerging microbial threats, and consider methods for their control. We emphasize that emergence results from dynamic interactions between rapidly evolving infectious agents and changes in the environment and in host behaviour that provide such agents with favourable new ecological niches.

Global burden of infectious diseases

About 15 million (>25%) of 57 million annual deaths worldwide are estimated to be related directly to infectious diseases; this figure does not include the additional millions of deaths that occur as a consequence of past infections (for example, streptococcal rheumatic heart disease), or because of complications associated with chronic infections, such as liver failure and hepatocellular carcinoma in people infected with hepatitis B or C viruses⁷ (Fig. 2).

The burden of morbidity (ill health) and mortality associated with infectious diseases falls most heavily on people in developing countries⁸, and particularly on infants and children (about three million children die each year from malaria and diarrhoeal diseases alone⁷). In developed nations, infectious disease mortality disproportionately affects indigenous and disadvantaged minorities⁹.

Emerging infections in historical context

EIs have been familiar threats since ancient times. They were once identified by terms such as *λοιμός* (*loimos*)¹⁰,

and later as ‘pestilences’, ‘pestes’, ‘pests’ and ‘plagues’. Many examples can be cited in addition to the Black Death and the 1918 influenza pandemic, such as certain biblical pharaonic plagues and the unidentified Plague of Athens, which heralded the end of Greece’s Golden Age¹¹. The Age of Discovery, starting in the fifteenth century, was a particularly disastrous period with regard to the spread of infectious diseases. Importation of smallpox into Mexico caused 10–15 million deaths in 1520–1521, effectively ending Aztec civilization^{12,13}. Other Amerindian and Pacific civilizations were destroyed by imported smallpox and measles^{13–17}. Historians have referred to these events as apocalypses¹⁶ and even as genocide¹⁵.

For centuries, mankind seemed helpless against these sudden epidemics. But the establishment of the germ theory and the identification of specific microbes as the causative agents of a wide variety of infectious diseases^{18–20} led to enormous progress, notably the development of vaccines and ultimately of antimicrobials²⁰. In fact, the era of the identification of microbes had barely begun¹⁸ when optimists at the end of the nineteenth century predicted the eradication of infectious diseases²¹. By the 1950s, which witnessed the widespread use of penicillin, the development of polio vaccines and the discovery of drugs for tuberculosis, complacency had set in²², and in 1967, the US Surgeon General stated that “the war against infectious diseases has been won”²³.

Some experts remained sceptical, aware of recurrent lessons from history. They were less persuaded by successes than alarmed by failures such as the lack of progress against infections in the developing world and the global spread of antimicrobial resistance. Richard Krause, then the director of the US National Institute of Allergy and Infectious Diseases, warned in 1981 (ref. 24) that microbial diversity and evolutionary vigour were still dynamic forces threatening mankind. As Krause was completing his book *The Restless Tide*²⁴, AIDS — one of history’s most devastating pandemics — was already insidiously emerging. The emergence of AIDS led to renewed appreciation of the inevitability and consequences of the emergence of infectious diseases^{25–31}. In the past 25 years, some of the factors that resulted in AIDS have also led to re-emergences of historically important diseases such as cholera, diphtheria, trench fever and plague. Many re-emergences have been catalysed by wars, loss of social cohesion, and natural disasters such as earthquakes and floods, indicating the importance not only of microbial and viral factors, but also of social and environmental determinants^{25–31}.

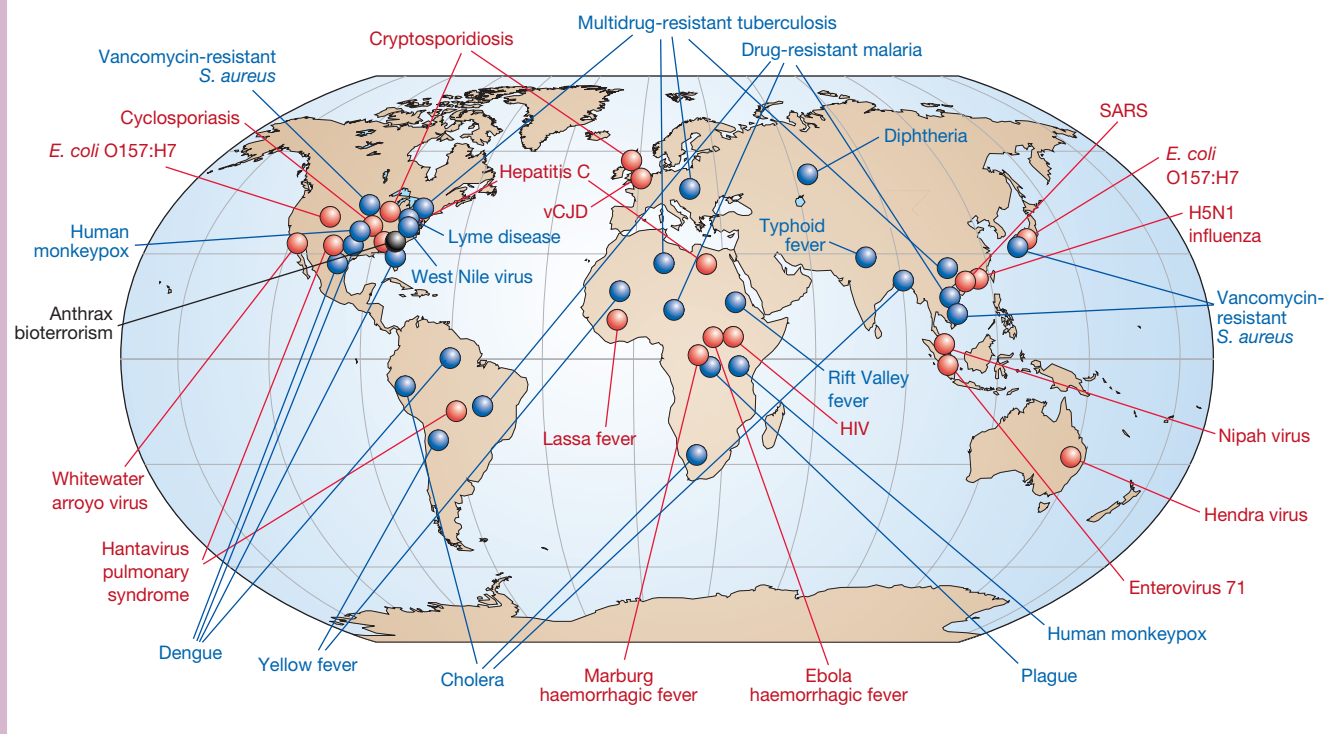


Figure 1 Global examples of emerging and re-emerging infectious diseases, some of which are discussed in the main text. Red represents newly emerging diseases; blue, re-emerging/resurging diseases; black, a 'deliberately emerging' disease. Adapted, with permission, from ref. 23.

Newly emerging and newly recognized infections

The classification of EIs as 'newly emerging', 're-emerging/resurging' or 'deliberately emerging' is useful because the underlying causes of emergence and the optimal prevention or control responses frequently differ between the groups. Newly emerging infections are those that have not previously been recognized in man. Many diverse factors contribute to their emergences (see Box 1); these include microbial genetic mutation and viral genetic recombination or reassortment, changes in populations of reservoir hosts or intermediate insect vectors, microbial switching from animal to human hosts, human behavioural changes (notably human movement and urbanization), and environmental factors. These numerous microbial, host and environmental factors interact to create opportunities for infectious agents to evolve into new ecological niches, reach and adapt to new hosts, and spread more easily between them.

The AIDS model

Any discussion of recent EIs must begin with the human immunodeficiency virus (HIV) that causes AIDS. HIV has so far infected more than 60 million people worldwide³³. Before jumping to humans an estimated 60–70 years ago³⁴, perhaps as a consequence of the consumption of 'bush meat' from non-human primates, HIV-1 and HIV-2 had ample opportunity to evolve in hosts that were genetically similar to man (the chimpanzee, *Pan troglodytes*, and the sooty mangabey, *Cercocebus atys*). But HIV/AIDS might never have emerged had it not been for disruptions in the economic and social infrastructure in post-colonial sub-Saharan Africa. Increased travel, the movement of rural populations to large cities, urban poverty and a weakening of family structure all promoted sexual practices, such as promiscuity and prostitution, that facilitate HIV transmission^{34–37}. Such complex interactions between infectious agents, hosts and the environment are not unique to the epidemiology of HIV/AIDS. The examples cited below further illustrate how changes in population density, human movements and the environment interact to create ecological niches that facilitate microbial or viral adaptation.

Dead-end transmission of zoonotic and vector-borne diseases

Some infectious agents that have adapted to non-human hosts can jump to humans but, unlike HIV, are not generally transmitted from person to person, achieving only 'dead end' transmission. Infections in animals that are transmitted to humans (zoonoses), and those transmitted from one vertebrate to another by an arthropod vector (vector-borne diseases), have repeatedly been identified as ranking among the most important EIs^{25,26}. Examples include the arenavirus haemorrhagic fevers (Argentine, Bolivian, Venezuelan and Lassa haemorrhagic fevers) and hantavirus pulmonary syndrome (HPS). Viruses in these groups have co-evolved with specific rodent species whose contact with humans has increased as a result of modern environmental and human behavioural factors. Farming, keeping domestic pets, hunting and camping, deforestation and other types of habitat destruction all create new opportunities for such infectious agents to invade human hosts^{25–31}. The first epidemic of HPS, detected in the southwestern region of the United States in 1993 (ref. 38), resulted from population booms of the deer mouse *Peromyscus maniculatis*, in turn caused by climate-related and recurrent proliferation of rodent food sources. Increased rodent populations and eventual shortages of food drove expanded deer mouse populations into homes, exposing people to virus-containing droppings. The 1998–1999 Malaysian Nipah virus epidemic³⁹ further illustrates the influence of human behaviours and environmental perturbations on newly emerging human infections. Pigs crammed together in pens located in or near orchards attracted fruit bats whose normal habitats had been destroyed by deforestation and whose droppings contained the then-unknown paramyxovirus. Virus aerosolization caused infection of pigs, with overcrowding leading to explosive transmission rates and ultimately to infections in pig handlers.

Variant Creutzfeldt–Jakob disease (vCJD) is another example of a zoonotic disease emerging in humans. vCJD is caused by the human-adapted form of the prion associated with the emerging epizootic (large-scale animal outbreak) of bovine spongiform encephalopathy (BSE)⁴⁰, commonly known as mad cow disease. The ongoing BSE

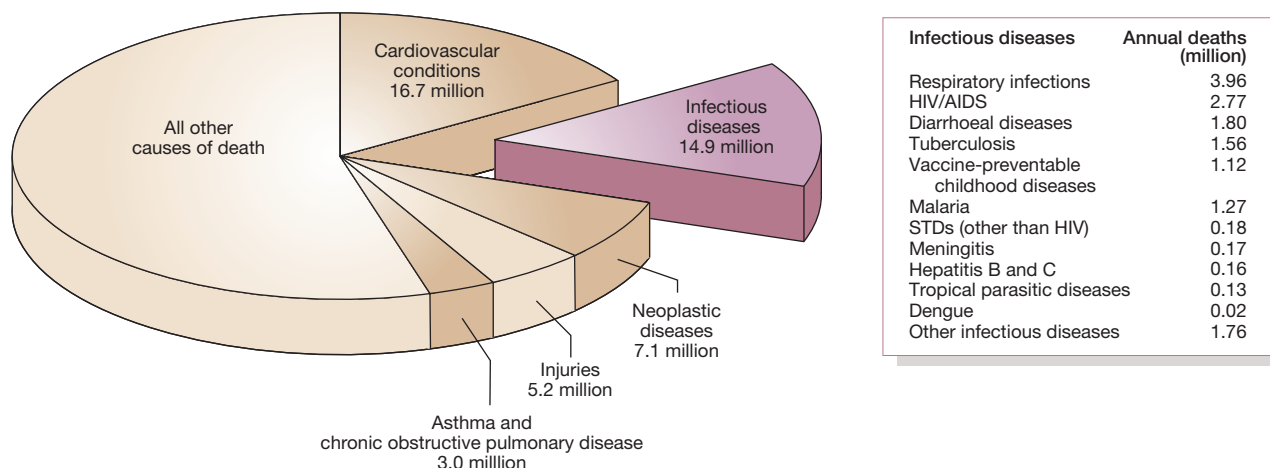


Figure 2 Leading causes of death worldwide. About 15 million (>25%) of 57 million annual deaths worldwide are the direct result of infectious disease. Figures published by the World Health Organization (see <http://www.who.int/whr/en> and ref. 7).

epizootic/vCJD epidemic, primarily affecting Great Britain, probably resulted from the now-abandoned practice of supplementing cattle feed with the pulverized meat and bones of previously slaughtered cattle. BSE itself is suspected to have emerged because of even earlier use of cattle feed containing the agent of sheep scrapie, a prion disease recognized by farmers more than 250 years ago⁴¹. Alarming, the new BSE prion has become uncharacteristically promiscuous: unlike most known prions, it readily infects multiple species in addition to humans. This suggests the possibility of further emerging diseases associated with prions with currently unknown transmissibility to humans⁴⁰. The recent reports of variant strains of the BSE prion⁴² suggest that the BSE agent could be a more serious threat than other animal prions.

Environmentally persistent organisms

Infectious agents indirectly transmitted to or between humans by way of human-modified environments account for other emerging zoonoses, as well as certain non-zoonotic diseases, which are discussed below. For example, legionnaires' disease, first identified in 1976, is caused by *Legionella pneumophila*, whose emergence as a human pathogen might not have occurred were it not for the environmental niche provided by air-conditioning systems²⁶. *Campylobacter jejuni* and Shiga-toxin-producing *Escherichia coli* (*E. coli* O157:H7 and other agents of haemolytic-uraemic syndrome) infect agricultural animals, gaining access to humans through food, milk, water or direct animal contact. Other enteric pathogens, such as the vibrios causing classical cholera (re-emerging; see below) and serogroup O139 cholera, and the zoonotic protozoa *Cryptosporidium parvum* and *Cyclospora cayentanensis*²⁶, seem to have come from environmental or animal organisms that have adapted to human-to-human 'faecal-oral' transmission through water.

Old microbes cause new diseases

Some EIs come from microorganisms that once caused familiar diseases, but which now cause new or previously uncommon diseases. *Streptococcus pyogenes* caused a fatal pandemic of scarlet and puerperal fevers between 1830 and 1900 (ref. 44). Scarlet fever, then the leading cause of death in children, is now rare, but has been largely supplemented by other streptococcal complications such as streptococcal toxic shock syndrome, necrotizing fasciitis and re-emergent rheumatic fever⁴⁵. When new microbes are discovered, their emergences as disease-causing pathogens may be delayed. For example, in 1883, Robert Koch was unable to show that the newly discovered

Koch-Weeks bacillus caused serious disease. More than a century later, a fatal EI dubbed Brazilian purpuric fever was linked to virulent clonal variants of *Haemophilus influenzae* biogroup *aegyptius* (the Koch-Weeks bacillus)⁴⁶. Although the bases of emergences of new and more severe diseases caused by *S. pyogenes* and *H. influenzae* biogroup *aegyptius* are not fully known, in both cases complex microbial genetic events are suspected. The distinctive clonal variants associated with severe *H. influenzae* biogroup *aegyptius* disease have been shown by PCR (polymerase chain reaction)-based subtractive genome hybridization to contain not only a unique plasmid, but also unique chromosomal regions, some of which are encoded by bacteriophages⁴⁷. This research has narrowed the search for virulence determinants to unique proteins, some of which may have been acquired from other organisms by horizontal gene transfer. *Streptococcus pyogenes* has been studied more extensively, but the basis of severe disease emergence seems to be more complex than for *H. influenzae* biogroup *aegyptius*. Many factors associated with streptococcal virulence have been identified in strains bearing the M1 surface protein as well as in other M protein strains, among them bacteriophage-encoded superantigen toxins and a protein known as sic (streptococcal inhibitor of complement), which seems to be strongly selected by human host mucosal factors. Several lines of evidence suggest that changes in streptococcal virulence reflect genetic changes associated with phage integration, large-scale chromosomal rearrangements and possibly the shuffling of virulence cassettes (clusters of genes responsible for pathogenicity), followed by rapid human spread and immune selection^{48,49}.

Microbial agents and chronic diseases

Infectious agents that are associated with chronic diseases are one of the most challenging categories of newly emerging (or at least newly appreciated) infections. Examples include the associations of hepatitis B and C with chronic liver damage and hepatocellular carcinoma, of certain genotypes of human papillomaviruses with cancer of the uterine cervix, of Epstein-Barr virus with Burkitt's lymphoma (largely in Africa) and nasopharyngeal carcinoma (in China), of human herpesvirus 8 with Kaposi sarcoma, and of *Helicobacter pylori* with gastric ulcers and gastric cancer⁵⁰⁻⁵². Some data even suggest infectious aetiologies for cardiovascular disease and diabetes mellitus⁵³, major causes of death and disability worldwide. Other associations between infectious agents and idiopathic chronic diseases will inevitably be found.

Re-emerging and resurging infections

Re-emerging and resurging infections are those that existed in the past but are now rapidly increasing either in incidence or in geographical or human host range. Re-emergence is caused by some of the factors that cause newly emerging infectious diseases, such as microbial evolutionary vigour, zoonotic encounters and environmental encroachment. Re-emergences or at least cyclical resurgences of some diseases may also be climate-related — for example, the El Niño/Southern Oscillation (ENSO) phenomenon is associated with resurgences of cholera and malaria⁵⁴.

Geographical spread of infections

The impact of both new and re-emerging infectious diseases on human populations is affected by the rate and degree to which they spread across geographical areas, depending on the movement of human hosts or of the vectors or reservoirs of infections. Travel has an important role in bringing people into contact with infectious agents⁵⁵. An increase in travel-associated importations of diseases was anticipated as early as 1933, when commercial air travel was still in its infancy⁵⁶. This has since been demonstrated dramatically by an international airline hub-to-hub pandemic spread of acute haemorrhagic conjunctivitis in 1981 (ref. 57), by epidemics of meningococcal meningitis associated with the Hajj, and more recently by the exportation of epidemic SARS (a newly emerging disease) from Guangdong Province, China, to Hong Kong, and from there to Beijing, Hanoi, Singapore, Toronto and elsewhere⁵ (Fig. 3). The persistent spread of HIV along air, trucking, drug-trafficking and troop-deployment routes is a deadly variation on this theme^{33–37}.

Malaria

Plasmodium falciparum malaria was neglected for several decades, but is now among the most important re-emerging diseases worldwide (Fig. 2). Years of effective use of dichlorodiphenyltrichloroethane (DDT) led to the abandonment of other mosquito-control programmes, but the insecticide fell into disuse because of mosquito resistance and concerns about the insecticide's potentially harmful effects on humans and wildlife. Consequently, malaria has re-emerged, and the situation has been worsened by the development of drug resistance to chloroquine and mefloquine⁵⁸. Research efforts focus on the development of vaccines⁵⁹ and new drugs, and on re-establishing public health measures such as the use of bed nets.

Tuberculosis

Tuberculosis is one of the most deadly re-emerging diseases (Fig. 2). The discovery of isoniazid and other drugs initially led to effective tuberculosis cures, empty sanatoria and the dismantling of public health control systems in developed nations. Consequently, by the 1980s, when tuberculosis had re-emerged in the era of HIV/AIDS, local and state health departments in the United States lacked field, laboratory and clinical staff and so had to reinvent tuberculosis-control programmes²⁵. The remarkable re-emergence of tuberculosis was fuelled by the immune deficiencies of people with AIDS, which greatly increases the risk of latent *Mycobacterium tuberculosis* infections progressing to active disease, and being transmitted to others. Inadequate courses of anti-tuberculosis therapy compound the problem, leading to the emergence and spread of drug-resistant and multidrug-resistant strains⁶⁰, and a need for more expensive treatment strategies such as directly observed therapy. It has been known for over a century that tuberculosis is a disease of poverty, associated with crowding and inadequate hygiene. The continuing expansion of global populations living in poverty makes tuberculosis more difficult to control.

Drug-resistant microbes

Drug resistance, another factor causing microbial and viral re-emergence, may result from mutation (for example, in the case of viruses and *M. tuberculosis*), or from bacterial acquisition of extraneous

Box 1

Factors involved in the emergence of infectious diseases

Selected factors contribute to the emergence/re-emergence of infectious diseases^{25,26}. These factors, which frequently differ for 'newly emerging', 're-emerging/resurging' and 'deliberately emerging' diseases, include genetic, biological, and social, political and economic factors.

- Microbial adaptation and change
- Human susceptibility to infection
- Climate and weather
- Changing ecosystems
- Human demographics and behaviour
- Economic development and land use
- International travel and commerce
- Technology and industry
- Breakdown of public health measures
- Poverty and social inequality
- War and famine
- Lack of political will
- Intent to harm

genes through transformation or infection with plasmids. Sequential emergences of *Staphylococcus aureus* that are resistant to sulpha drugs (1940s), penicillin (1950s), methicillin (1980s) and to vancomycin in 2002 (ref. 61) — a last line of antibiotic defence for some multiply drug-resistant bacteria — are troubling. Nosocomial *Enterococcus faecalis* became fully resistant to vancomycin by 1988, and then apparently transferred *vanA* resistance genes to co-infecting staphylococci⁶¹. Methicillin-resistant staphylococci are now being isolated from livestock that have been fed with growth-promoting antibiotics⁶², possibly contributing to resistance problems in humans. Many other important microbes have also become effective 'resistors', among them *Streptococcus pneumoniae* and *Neisseria gonorrhoeae*⁶³.

Opportunistic re-emerging infections

Immune deficiency associated with AIDS, and with chemotherapy for cancer, immune-mediated diseases and transplantation, has contributed to an enormous global increase in the numbers of immunosuppressed people over the past few decades (probably more than 1% of the world's population), setting the stage for the re-emergence of many opportunistic infections. HIV, which has infected more than 60 million people globally³³, is the largest single cause of human immune deficiency and markedly increases vulnerability to a wide range of opportunistic pathogens, including *Pneumocystis carinii*, various fungi, tuberculosis, protozoa and herpesviruses⁶⁴. Breakthroughs in cancer therapy and in immunosuppressive therapies used to treat immune-mediated diseases and for transplantation^{65,66} can also leave patients susceptible to opportunistic infections. Human organ transplantation adds a further risk of infection with undetected pathogens in donor tissues, and transplantation of animal organs introduces the risk of transmission to humans of animal microbes⁶⁷.

Re-emerging zoonotic and vector-borne diseases

The emergence of zoonotic and vector-borne diseases can also be associated with human behaviours and environmental perturbation. In 2003, monkeypox — an endemic infection of African rodents — crossed the Atlantic with exported pets, which were then shipped from Texas to infect people throughout the US Midwest⁴. Lyme disease, caused by *Borrelia burgdorferi*, re-emerged in the United States as a result of suburban expansion, which brought people into increasing contact with deer, deer mice and ticks. Similarly, tick-borne encephalitis re-emerged in Russia when weekend getaways

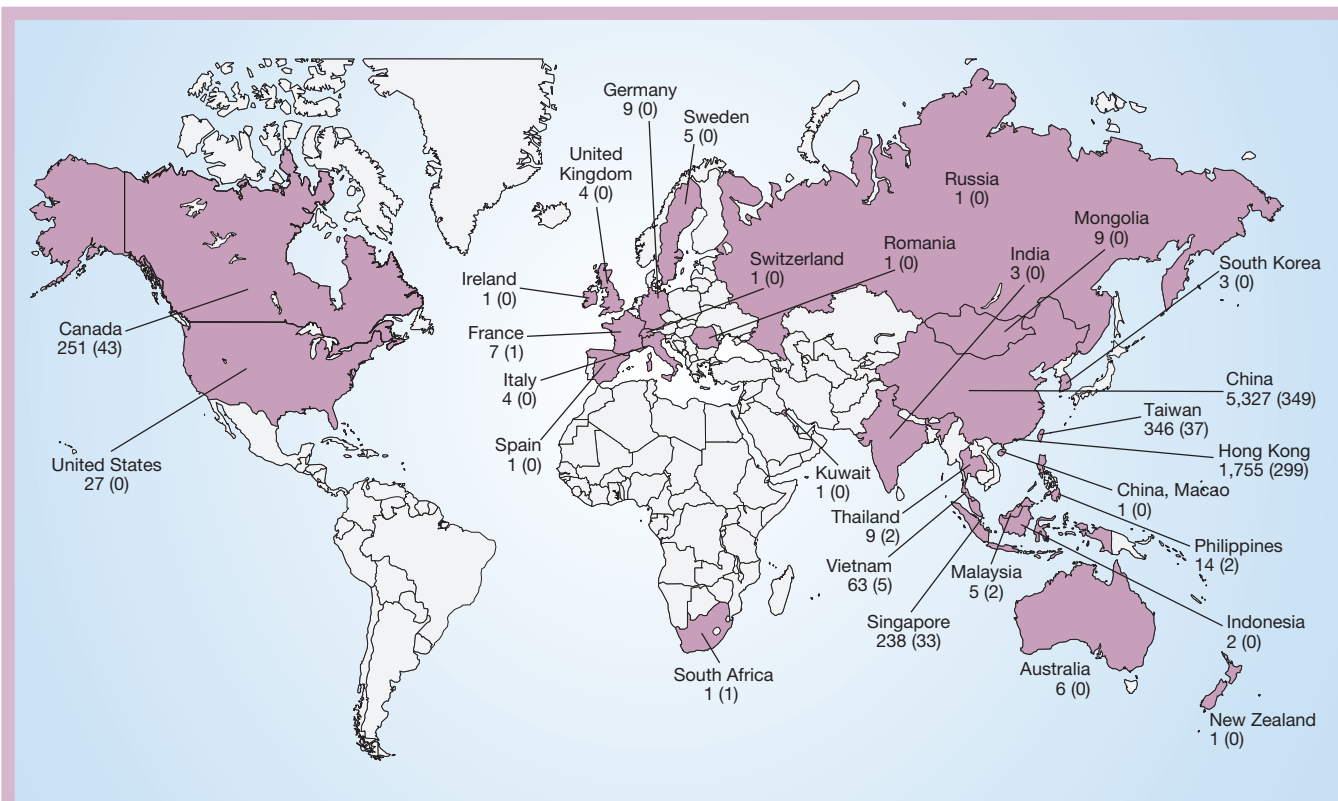


Figure 3 Probable cases of severe acute respiratory syndrome (SARS) with onset of illness from 1 November 2002 to 31 July 2003. Cases are given by country. SARS-related deaths are indicated in parentheses. A total of 8,096 cases (and 774 deaths) are presented. Figures published by the World Health Organization (see <http://www.who.int/csr/sars/country/en>).

(dachas) drew city dwellers into contact with forest ticks. The simultaneous 1999 emergences of encephalitis due to West Nile virus in the United States and in Russia^{68,69} reflect abundances of eclectic vector mosquitoes and avian hosts in these locations. Both were probably connected to endemic sites by virus carriage in migratory birds and travellers. The remarkable geographical spread of West Nile virus in the five years since its introduction into the Western Hemisphere reflects an unfortunate confluence of viral promiscuity and ecological diversity⁷⁰. Although humans are dead-end hosts for West Nile virus, the risk of infection is greatly increased by marked zoonotic viral amplification and persistence in the environment. Unlike most viruses, which tend to be fairly narrowly adapted to specific hosts, West Nile virus is known to infect more than 30 North American mosquito species, which together transmit infection to at least 150 North American bird species, many of which migrate to new and distant locations, spreading the virus to rural and urban ecosystems throughout North and Central America⁷⁰.

Although West Nile virus is now a major epidemiological concern in the developed world, dengue remains the most significant and widespread flavivirus disease to have emerged globally⁷¹. A 2001–2002 epidemic in Hawaii — fortunately without fatalities — is a reminder that dengue has also re-emerged in locations once considered to be dengue-free. Usually transmitted by *Aedes aegypti* mosquitoes, dengue has recently been transmitted by *Aedes albopictus* — a vector switch of potential significance with respect to dengue re-emergence⁷¹. In the Americas, including many US southern states, *A. albopictus* has been spreading into areas where *A. aegypti* mosquitoes are not found, and persisting for longer seasonal periods, putting tens of millions more people at risk of dengue infection. Dengue re-emergence is further complicated by disturbing increases in a serious and formerly rare form of the disease, dengue haemorrhagic fever (dengue shock syndrome being its highly fatal form). These severe complications are thought to result from the evolution of dengue viruses to escape high

population immunity, seen in increased viral virulence and human immunopathogenesis due to antibody-dependent enhancement of viral infection⁷².

Cholera is also of interest, not only as an important cause of mortality, but also because of the complexity of factors that determine its re-emergence. Both virulent and avirulent strains of these zoonotic bacteria are maintained in the environment and are rapidly evolving in association with phyto- and zooplankton, algae and crustaceans. Such environmental strains seem to act as reservoirs for human virulence genes (for example, genes for the phage-encoded cholera toxin and the toxin-coregulated pilus (TCP) factor associated with attachment), and to undergo gene transfer events that lead to new strains containing further virulence gene combinations. These result in periodic cholera emergences that cause epidemics and pandemics⁷³. Thus, although the disease we know as cholera has appeared to be clinically and epidemiologically stable at least since the third pandemic (in the 1840s), modern evidence suggests that such apparent stability masks aggressive bacterial evolution in complex natural environments.

Influenza A

Influenza A viruses, which are endemic gastrointestinal viruses of wild waterfowl, have evolved elaborate mechanisms to jump species into domestic fowl, farm animals and humans. Periodic gene segment reassortments between human and animal viruses produce important antigenic changes, referred to as 'shifts'. These can lead to deadly pandemics, as occurred in 1888, 1918, 1957 and 1968 (refs 74, 75). In intervening years, shifted viruses undergo continual but less dramatic antigenic changes called 'drifts', which allow them partially to escape human immunity raised by previously circulating influenza viruses. Influenza drift is an evolutionary success story for the virus. Influenza A has a seemingly inexhaustible repertoire of mutational possibilities at several critical epitopes surrounding the

viral haemagglutinin site that attaches to human cells. It remains something of a mystery how zoonotic influenza viruses mix with each other and with human strains to acquire the additional properties of human virulence and human-to-human transmissibility.

Before 1997, mild cases of human disease associated with avian influenza viruses were occasionally reported⁷⁶. These events have become more frequent, sometimes resulting in severe cases of disease and death. Avian influenza has recently made dead-end jumps to humans — for example, the 1997 Hong Kong outbreak of the newly emergent H5N1 influenza, the 2003 H7N7 epidemic in the Netherlands, the 2003–2004 H5N1 and H7N3 epizootics in Asia and elsewhere, and occasional cases of H9N2 disease (Fig. 4). Meanwhile, back-switches of human H3N2 viruses have emerged in pigs, from which both doubly mixed (pig–human) and triply mixed (pig–human–avian) viruses^{74,75} have been isolated. Such enzootic/zoonotic mixing is suspected to have occurred in the influenza pandemic of 1918–1920, which was caused by an H1N1 virus with an avian-like receptor-binding site⁷⁷. The predicted virulence genes of this virus are now being sought from 85-year-old pathology specimens and from frozen corpses⁷⁸. The implications of interspecies genetic mixing for future influenza pandemics are troubling. Although much remains speculative about how influenza viruses emerge and spread, it seems clear that the process is driven by prolific and complex viral evolution (genetic reassortment and mutational ‘drift’), interspecies mixing and adaptation, and ecological factors that bring humans into contact with animals and each other. By whatever means new influenza virus pandemic strains emerge, they eventually reach a critical threshold of human transmission beyond which epidemic and pandemic spread follows mathematically predictable patterns.

The dynamics and determinants of such epidemic development have been studied since the nineteenth century for several infectious diseases. For influenza, both historical and prospective epidemics have been described or predicted using deterministic and stochastic mathematical models, often with surprising accuracy when compared with actual epidemic data. More complicated mathematical models that describe how diseases spread by means other than person-to-person aerosol transmission have generally been less successful in describing and predicting epidemics, but have nonetheless been helpful in planning public health responses to epidemics caused by HIV⁷⁹, vCJD⁸⁰ and other diseases.

Mathematical modelling is also used to determine the impact of emerging epidemics. For example, it has been difficult to estimate overall influenza mortality because fatal infections are often neither diagnosed nor accurately recorded in hospital records and death certificates, especially in the elderly. Recent epidemiological attempts to obtain improved influenza mortality estimates from seasonal excess mortality data⁸¹ have indicated that influenza mortality may be

greater than was previously suspected, because influenza deaths are frequently coded under seemingly unrelated categories such as cardiovascular diseases. The same approaches also show that other influenza-like deaths may actually be due to other agents, such as respiratory syncytial virus (RSV), a common childhood virus that in the past decade has emerged as a major cause of adult mortality⁸¹.

Deliberately emerging infections

Deliberately emerging microbes are those that have been developed by man, usually for nefarious use. The term ‘deliberately emerging’ refers to both naturally occurring microbial agents such as anthrax⁶, and to bioengineered microorganisms such as those created by the insertion of genetic virulence factors that produce or exacerbate disease. Deliberately emerging microbes include microorganisms or toxins produced in a form that would cause maximal harm because of ease of dissemination, enhanced infectivity or heightened pathogenicity⁸².

Bioterrorism and biowarfare

As concepts, bioterrorism and biowarfare are probably not new. The alleged catapulting of plague-ridden corpses over enemy walls in the 1346 siege of Caffa (the modern Crimean port of Feodosia, Ukraine) and the dispatch of smallpox-impregnated blankets to Indians by British officers in the Seven Years War (1754–1763) have frequently been cited as examples of bioterrorism or biowarfare^{83,84}.

Two modern attacks have been well documented. In 1984, an Oregon religious cult spiked restaurant salad bars with *Salmonella* in an attempt to sway a local election⁸⁵. A 2001 anthrax attack⁶, in which a terrorist mailed anthrax-spore-filled letters to prominent figures, including two US senators, resulted in illness in at least 18 people and the death of five of these individuals. Public alarm was elevated by the knowledge that *Bacillus anthracis* is a common and easily obtainable enzootic and soil organism found in laboratories worldwide, and that scientific technology had increased its lethality: the spores had been weaponized by being concentrated, finely milled and packed with a dispersal agent to increase their capacity to disseminate⁸². The United States, the United Kingdom, the Soviet Union and other nations once had sophisticated offensive bioweapons programmes that included the production of weaponized anthrax spores⁸². Soviet scientists continued to produce large quantities of organisms adapted for biowarfare and bioterror — among them the agents of smallpox, plague, tularaemia and Marburg virus — for several years after their signing of the Bioweapons and Toxins Treaty Convention in 1972, which forbade such activities⁸². By 1987, the Soviet programme was annually producing 5,000 tonnes of weaponized anthrax spores, packing them into warheads and other delivery devices⁸².

Before the 2001 anthrax attacks⁶, the US scientific community had for several years been bolstering its biodefence research capacity. The anthrax attacks greatly accelerated this expansion as part of a national defence plan, which includes efforts to provide a knowledge base for the development of effective countermeasures against agents of bioterror, such as diagnostics, therapeutics and vaccines, and to translate this knowledge into the production and delivery of such measures⁸⁶. Bioterror agents have been grouped into three categories of risk⁸⁷. The six category A agents (anthrax, smallpox, plague, tularaemia, viral haemorrhagic fevers and clostridial botulinum toxin) are given top priority because they are highly lethal and readily deployed as weapons. Category B and C agents include food-borne and water-borne organisms that incapacitate but usually do not kill.

Meeting the challenge of emerging infections

Infectious diseases will continue to emerge and re-emerge, leading to unpredictable epidemics and difficult challenges to public health and to microbiology and allied sciences.

Surveillance and response, the key elements in controlling EIs, be they naturally occurring or deliberately engineered, depend on rapid clinical diagnosis and detection and containment in populations and

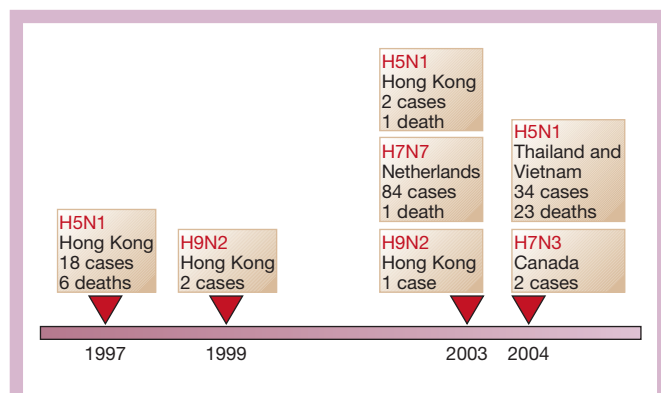


Figure 4 Documented human infections with avian influenza viruses, 1997–2004. Sporadic cases of mild human illness associated with avian influenza viruses were reported before 1997. See <http://www.who.int/csr/disease/influenza/en> and ref. 76.

in the environment. Globally, such efforts are coordinated by the World Health Organization, which recently led a multifaceted effort to successfully contain the global SARS outbreak of 2002–2003 (ref. 88). In the United States, such efforts are led by the US Centers for Disease Control and Prevention (CDC)⁸⁹, which along with state and local health departments and other agencies have been making significant strides in national surveillance–response capacity. The enormous influx of US government-funded research resources (largely through the National Institutes of Health) and public health resources (mainly through the CDC, and state and local public health agencies) in response to the increased threat of a bioterrorist attack⁸⁶ will fortify the response capabilities related to all EIs.

However, it is clear that surveillance and other activities that traditionally fall within the domain of public health are not in themselves sufficient to adequately address the problem of EIs. Of critical importance are basic, translational and applied research efforts to develop advanced countermeasures such as surveillance tools, diagnostic tests, vaccines and therapeutics⁸⁶. Genomics, proteomics and advances in nanotechnology⁹⁰ are increasingly being exploited in diagnostic, therapeutic and microbial research applications, and in rational drug and vaccine design. Direct and computational structural determination⁹¹, prediction of protein–protein interactions between microorganisms and drugs, and sophisticated bioinformatics techniques support research in all of the above areas. These technologies have led to numerous advances in real-world utility against EIs, most notably in the development of more than 20 antiretroviral drugs that can effectively suppress HIV replication. Where they are available and properly used in HIV-infected individuals, these medications have dramatically reduced HIV morbidity and mortality⁹².

Gene- and protein-based microarrays can be used to detect pathogen signals, to monitor resistance to anti-infective agents, to characterize host gene responses to recent infections, and to facilitate the development of new drugs and vaccines⁹³. Basic and applied research together have provided promising new vaccine platforms, such as recombinant proteins, immunogenic peptides, naked DNA vaccines, viral vectors of extraneous genes encoding immunogenic proteins (including chimaeras), replicons and pseudovirions⁹⁴. Many novel vaccine candidates are now being developed against EIs such as HIV, Ebola virus, West Nile virus, dengue, the SARS coronavirus, tuberculosis and malaria. Of particular note are novel tuberculosis vaccines that recently entered clinical trials — the first time in more than 60 years that new approaches to vaccination for tuberculosis have been assessed in humans⁹⁵. Chimaeric flavivirus vaccines for West Nile virus, dengue and Japanese encephalitis virus are effective in animal models and are in various stages of clinical testing⁹⁵.

Our growing understanding of the human immune system is also helping to accelerate vaccine development. This is especially true in the case of innate immune responses, which are evolutionarily older, less specific and faster-acting than the adaptive responses that have been the traditional targets of vaccines⁹⁶. As we learn more about innate immunity and its relationship with the adaptive immune system, opportunities to create more effective vaccine adjuvants will emerge. For example, synthetic DNA sequences that contain repeated CpG motifs mimic the stimulatory activity that bacterial DNA fragments exert on the innate immune system. These sequences show promise as vaccine adjuvants that accelerate and augment immune responses⁹⁷. We can anticipate more progress of this kind as we continue to delineate the complex interactions between innate and adaptive immune responses.

The sequencing of the human genome, the genomes of six other animals, including the mouse, and those of microbial vectors and microbes themselves (for example, *P. falciparum* and its mosquito vector, *Anopheles gambiae*), have elevated microbiology to a whole-genome level. The ability to sequence microbial genomes in a few days⁹⁸ or less, and to examine host–vector–microbe interactions at both the genome level and at the tertiary protein structural level, will help us to understand the molecular mechanisms that underlie the

pathogenesis of infectious disease and host defences, including resistance and immune evasion. These advances will facilitate the development of new countermeasures. Other fertile areas of research include the use of geographical information systems⁹⁹ and satellite imaging to support field study and epidemic prevention (for example, predicting HPS and Rift Valley fever epidemics in indigenous areas by satellite imagery of water and vegetation related to animal reservoir and vector prevalence).

Underlying disease emergence are evolutionary conflicts between rapidly evolving and adapting infectious agents and their slowly evolving hosts. These are fought out in the context of accelerating environmental and human behavioural alterations that provide new ecological niches into which evolving microbes can readily fit. It is essential that broadly based prevention strategies, as well as new and improved countermeasures (that is, surveillance tools, diagnostics, therapeutics and vaccines), be continually tested, refined and upgraded, requiring a strengthened relationship between public health and basic and clinical sciences. The challenge presented by the ongoing conflict between pathogenic microorganisms and man has been well summarized by a noted champion of the war on EIs, Joshua Lederberg: “The future of microbes and mankind will probably unfold as episodes of a suspense thriller that could be entitled *Our Wits Versus Their Genes*”²⁹. The global scientific and public health communities must confront this reality not only with wit, but also with vision and sustained commitment to meet a perpetual challenge. □

doi:10.1038/nature02759

- Morse, S. S. Factors in the emergence of infectious diseases. *Emerg. Infect. Dis.* **1**, 7–15 (1995).
- Kohn, G. C. in *The Wordsworth Encyclopedia of Plague and Pestilence* (ed. Kohn, G. C.) 25–26 (Facts on File, New York, 1995).
- Johnson, N. P. A. S. & Mueller, J. Updating the accounts: global mortality of the 1918–1920 “Spanish” influenza pandemic. *Bull. Hist. Med.* **76**, 105–115 (2002).
- Centers for Disease Control and Prevention. Multistate outbreak of monkeypox — Illinois, Indiana, and Wisconsin, 2003. *MMWR* **52**, 537–540 (2003).
- Peiris, J. S. M. *et al.* The severe acute respiratory syndrome. *N. Engl. J. Med.* **349**, 2431–2441 (2003).
- Jernigan, D. B. *et al.* Investigation of bioterrorism-related anthrax, United States, 2001: epidemiologic findings. *Emerg. Infect. Dis.* **8**, 1019–1028 (2002).
- World Health Organization. *The World Health Report 2004* (World Health Organization, Genève, 2004).
- Guerrant, R. L. & Blackwood, B. L. Threats to global health and survival: the growing crises of tropical infectious diseases — an “unfinished” agenda. *Clin. Infect. Dis.* **28**, 966–986 (1999).
- Butler, J. C. *et al.* Emerging infectious diseases among indigenous peoples. *Emerg. Infect. Dis.* **7** (suppl.), 554–555 (2001).
- Hodges, N. Δομολογία — sive, *Pestis Nuperae apud Populum Londinensem Grassantis Narratio Historica* (Josephi Nevill, London, 1672).
- Morens, D. M. & Littman, R. J. Epidemiology of the Plague of Athens. *Trans. Am. Philol. Assoc.* **122**, 271–304 (1992).
- Crosby, A. W. in *The Columbian Exchange: Biological and Cultural Consequences of 1492* (Greenwood, Westport, Connecticut, 1972).
- Hopkins, D. in *Princes and Peasants. Smallpox in History* Ch. 6, 204–233 (Univ. Chicago Press, Chicago, 1983).
- MacArthur, N. *Island Populations of the Pacific* (Australian National Univ. Press, Canberra, 1968).
- Bushnell, O. A. *The Gifts of Civilization. Germs and Genocide in Hawaii* (Univ. Hawaii Press, Honolulu, 1993).
- De Bevoise, K. *Agents of Apocalypse: Epidemic Disease in the Colonial Philippines* (Princeton Univ. Press, Princeton, New Jersey, 1995).
- Morens, D. M. Measles in Fiji, 1875: thoughts on the history of emerging infectious diseases. *Pac. Health Dialog* **5**, 119–128 (1998).
- Koch, R. Untersuchungen über Bakterien. V. Die Aetiologie der Milzbrand-Krankheit, begründet auf die Entwicklungsgeschichte des Bacillus Anthracis. *Beiträge zur Biologie der Pflanzen* **2**, 277–310 (1876).
- Worboys, M. *Spreading Germs: Diseases, Theories, and Medical Practice in Britain, 1865–1900* (Cambridge Univ. Press, Cambridge, 2000).
- Porter, R. *The Greatest Benefit to Mankind: A Medical History of Humanity from Antiquity to the Present* (W. W. Norton and Co., London, 1997).
- Deming, W. C. The extermination of infectious diseases. *NY Med. J.* **59**, 710–715 (1894).
- Cockburn, A. *The Evolution and Eradication of Infectious Diseases* (Johns Hopkins Univ. Press, Baltimore, 1963).
- Fauci, A. S. Infectious diseases: considerations for the 21st century. *Clin. Infect. Dis.* **32**, 675–685 (2001).
- Krause, R. M. *The Restless Tide: The Persistent Challenge of the Microbial World* (National Foundation for Infectious Diseases, Washington DC, 1981).
- Committee on Emerging Microbial Threats to Health. *Emerging Infections. Microbial Threats to Health in the United States* (eds Lederberg, J., Shope, R. E. & Oaks, S. C.) (National Academy Press, Washington DC, 1992).
- Committee on Emerging Microbial Threats to Health in the 21st Century. *Microbial Threats to Health in the United States: Emergence, Detection and Response* (eds Smolinski, M. S., Hamburg, M. A. & Lederberg, J.) (National Academy Press, Washington DC, 2003).

27. Stephens, D. S. *et al.* Emerging and re-emerging infectious diseases: a multidisciplinary perspective. *Am. J. Med. Sci.* **315**, 64–75 (1998).
28. Desselberger, U. Emerging and re-emerging infectious diseases. *J. Infect.* **40**, 3–15 (2000).
29. Lederberg, J. Infectious history. *Science* **288**, 287–293 (2000).
30. Pollard, A. J. & Dobson, A. R. Emerging infectious diseases in the 21st century. *Curr. Opin. Infect. Dis.* **13**, 265–275 (2000).
31. Feldman, H. *et al.* Emerging and re-emerging infectious diseases. *Med. Microbiol. Immunol.* **191**, 63–74 (2002).
32. McMichael, T. *Human Frontiers, Environments and Disease. Past Patterns, Uncertain Futures* (Cambridge Univ. Press, Cambridge, 2001).
33. Joint United Nations Programme on HIV/AIDS. *AIDS Epidemic Update: December 2003* (UNAIDS, Genève, 2003).
34. Sharp, P. M. *et al.* The origins of acquired immune deficiency viruses: where and when? *Phil. Trans. R. Soc. Lond. B* **356**, 867–876 (2001).
35. Quinn, T. C. Population migration and the spread of types 1 and 2 human immunodeficiency viruses. *Proc. Natl Acad. Sci. USA* **91**, 2407–2414 (1994).
36. Daley, C. L. *et al.* in *The AIDS Knowledge Base* 3rd edn (eds Cohen, P. T., Sande, M. A. & Volberding, P. A.) Ch. 3, 23–52 (Lippincott, Williams and Wilkins, Philadelphia, 1999).
37. Ronald, A. R. Slowing heterosexual HIV transmission. *Infect. Dis. Clin. North Am.* **9**, 287–296 (1995).
38. Hjelle, B. *et al.* A novel hantavirus associated with an outbreak of fatal respiratory disease in the southwestern United States: evolutionary relationships to known hantaviruses. *J. Virol.* **168**, 592–596 (1994).
39. Chua, K. B. *et al.* Nipah virus: a recently emergent deadly paramyxovirus. *Science* **288**, 1432–1435 (2000).
40. Beisel, C. D. & Morens, D. M. Variant Creutzfeldt–Jakob disease and the acquired and transmissible spongiform encephalopathies. *Clin. Infect. Dis.* **38**, 697–704 (2004).
41. Leopoldt, J. G. *Nützliche und auf die Erfahrung gegründete Einleitung zu der Landwirtschaft, fünf Theile* 348 (Christian Friedrich Günther, Berlin, 1759).
42. Casalone, C. *et al.* Identification of a second bovine amyloidotic spongiform encephalopathy: molecular similarities with sporadic Creutzfeldt–Jakob disease. *Proc. Natl Acad. Sci. USA* **101**, 3065–3070 (2004).
43. Williams, E. S. Scrapie and chronic wasting disease. *Clin. Lab. Med.* **23**, 139–159 (2003).
44. Katz, A. R. & Morens, D. M. Severe streptococcal infections in historical perspective. *Clin. Infect. Dis.* **14**, 298–307 (1992).
45. Musser, J. M. & Krause, R. M. in *Emerging Infections. Biomedical Research Reports* (ed. Krause, R. M.) 185–218 (Academic Press, San Diego, 1998).
46. Musser, J. M. & Selander, R. K. Brazilian purpuric fever: evolutionary genetic relationships of the case clone of *Haemophilus influenzae* biogroup *aegyptius* to encapsulated strains of *Haemophilus influenzae*. *J. Infect. Dis.* **161**, 130–133 (1990).
47. Li, M.-S., Farrant, J. L., Langford, P. R. & Kroll, J. S. Identification and characterization of genomic loci unique to the Brazilian purpuric fever clonal group of *H. influenzae* biogroup *aegyptius*: functionality explored using meningococcal homology. *Mol. Microbiol.* **47**, 1101–1111 (2003).
48. Reid, S. D., Hoe, N. P., Smoot, L. M. & Musser, J. M. Group A streptococcus: allelic variation, population genetics, and host–pathogen interactions. *J. Clin. Invest.* **107**, 393–399 (2001).
49. Beres, S. B. *et al.* Genome sequence of a serotype M3 strain of group A *Streptococcus*: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc. Natl Acad. Sci. USA* **99**, 10078–10083 (2003).
50. Chang, Y. *et al.* Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* **266**, 1865–1869 (1994).
51. Parsonnet, J. (ed.) *Microbes and Malignancy: Infection as a Cause of Human Cancers* (Oxford Univ. Press, New York, 1999).
52. Sanders, M. K. & Peura, D. A. *Helicobacter pylori*-associated diseases. *Curr. Gastroenterol. Rep.* **4**, 448–454 (2002).
53. Fredricks, D. N. & Relman, D. A. in *Current Clinical Topics in Infectious Diseases* Vol. 18 (eds Remington, J. S. & Swartz, M. N.) 180–200 (Blackwell Science, Malden, Massachusetts, 1998).
54. Kovats, R. S., Bouma, M. J., Hajat, S., Worrall, E. & Haines, A. El Niño and health. *Lancet* **362**, 1481–1489 (2003).
55. Cliff, A., Haggett, P. & Smallman-Raynor, M. in *Island Epidemics* Ch. 6, 165–236 (Oxford Univ. Press, Oxford, 2000).
56. Massey, A. *Epidemiology in Relation to Air Travel* (H. K. Lewis and Co., London, 1933).
57. Morens, D. M. Acute haemorrhagic conjunctivitis: dealing with a newly emerging disease. *Pac. Health Dialog* **5**, 147–153 (1998).
58. Welles, T. E. & Miller, L. H. Two worlds of malaria. *N. Engl. J. Med.* **349**, 1496–1498 (2003).
59. Miller, L. H. & Hoffman, S. L. Research toward vaccines against malaria. *Nature Med.* **4** (suppl.), 520–524 (1998).
60. Espinal, M. A. The global situation of MDR-TB. *Tuberculosis* **83**, 44–51 (2003).
61. Centers for Disease Control and Prevention. *Staphylococcus aureus* resistant to vancomycin — United States, 2002. *MMWR* **51**, 565–567 (2002).
62. Lee, J. H. Methicillin (oxacillin)-resistant *Staphylococcus aureus* strains isolated from major food animals and their potential transmission to humans. *Appl. Environ. Microbiol.* **69**, 6489–6494 (2003).
63. Neu, H. C. The crisis in antibiotic resistance. *Science* **257**, 1064–1072 (1992).
64. Jones, J. L. *et al.* Surveillance for AIDS-defining opportunistic illnesses, 1992–1997. *MMWR* **48** (CDC Surveillance Summary no. SS-2), 1–22 (1999).
65. Vento, S. & Cainelli, F. Infections in patients with cancer undergoing chemotherapy: aetiology, prevention, and treatment. *Lancet Oncol.* **4**, 595–604 (2003).
66. Singh, N. Impact of current transplantation practices on the changing epidemiology of infections in transplant recipients. *Lancet Infect. Dis.* **3**, 156–161 (2003).
67. Chapman, L. E. Xenotransplantation: public health risks — patient vs. society in an emerging field. *Curr. Top. Microbiol. Immunol.* **278**, 23–45 (2003).
68. Nash, D. *et al.* The outbreak of West Nile virus infection in the New York City area in 1999. *N. Engl. J. Med.* **344**, 1807–1814 (2001).
69. Platonov, A. E. *et al.* Outbreak of West Nile infection, Volgograd Region, Russia, 1999. *Emerg. Infect. Dis.* **7**, 128–132 (2001).
70. Komar, N. West Nile virus: epidemiology and ecology in North America. *Adv. Virus Res.* **61**, 185–234 (2003).
71. Gubler, D. J. Dengue and dengue haemorrhagic fever. *Clin. Microbiol. Rev.* **11**, 480–497 (1998).
72. Morens, D. M. Antibody-dependent enhancement of infection and the pathogenesis of viral disease. *Clin. Infect. Dis.* **19**, 500–512 (1994).
73. Faruque, S. M. & Nair, G. B. Molecular ecology of toxigenic *Vibrio cholerae*. *Microbiol. Immunol.* **46**, 59–66 (2002).
74. Shortridge, K. F., Peiris, J. S. & Guan, Y. The next influenza pandemic: lessons from Hong Kong. *J. Appl. Microbiol.* **94** (suppl.), 70S–79S (2003).
75. Webster, R. G. A molecular whodunit. *Science* **293**, 1773–1775 (2001).
76. Subbarao, K. & Katz, J. Avian influenza viruses infecting humans. *Cell Mol. Life Sci.* **571**, 1770–1784 (2000).
77. Stevens, J. *et al.* Structure of the uncleaved human H1 haemagglutinin from the extinct 1918 influenza virus. *Science* **303**, 1866–1870 (2004).
78. Reid, A. H. & Taubenberger, J. K. The origin of the 1918 pandemic influenza virus: a continuing enigma. *J. Gen. Virol.* **84**, 2285–2292 (2003).
79. Anderson, R. M., Swinton, J. & Garnett, G. P. Potential impact of low-efficacy HIV-1 vaccines in populations with high rates of infection. *Proc. R. Soc. Lond. B* **261**, 147–161 (1995).
80. Ghani, A. C., Ferguson, N. M., Donnelly, C. A. & Anderson, R. M. Short-term projections for variant Creutzfeldt–Jakob disease onsets. *Stat. Methods Med. Res.* **12**, 191–201 (2003).
81. Thompson, W. W. *et al.* Mortality associated with influenza and respiratory syncytial virus in the United States. *J. Am. Med. Assoc.* **289**, 179–186 (2003).
82. Alibek, K. & Handelman, S. *Biohazard: The Chilling True Story of the Largest Covert Biological Weapons Program in the World — Told From the Inside by the Man Who Ran It* (Random House, New York, 1999).
83. Henschel, A. W. Document zur Geschichte des schwarzen Todes. Mitgeteilt und eingeleitet. *Archiv für die gesamte Medizin* **2**, 26–59 (1842).
84. Duffy, J. Smallpox and the Indians in the American colonies. *Bull. Hist. Med.* **25**, 324–341 (1951).
85. Török, T. *et al.* A large community outbreak of salmonellosis caused by intentional contamination of restaurant salad bars. *J. Am. Med. Assoc.* **278**, 389–395 (1997).
86. Fauci, A. S. Biodefence on the research agenda: the world needs new and creative ways to counter bioterrorism. *Nature* **421**, 787 (2003).
87. Darling, R. G., Catlett, C. L., Huebner, K. D. & Jarrett, D. G. Threats in bioterrorism: I. CDC category A agents. *Emerg. Med. Clin. North Am.* **20**, 273–309 (2002).
88. World Health Organization. *Consensus Document on the Epidemiology of Severe Acute Respiratory Syndrome (SARS)* <<http://www.who.int/entity/csr/sars/en/WHOconsensus.pdf>> (2003).
89. Centers for Disease Control and Prevention. *Preventing Emerging Infectious Diseases: A Strategy for the 21st Century* (Department of Health and Human Services, Atlanta, 1998).
90. NIH Bioengineering Consortium. *Nanoscience and Nanotechnology: Shaping Biomedical Research*. Symposium Report <http://grants.nih.gov/grants/becon/becon_symposia.htm> (2002).
91. Baker, D. & Salí, A. Protein structure prediction and structural genomics. *Science* **294**, 93–96 (2001).
92. Dybul, M. *et al.* Guidelines for using antiretroviral agents among HIV-infected adults and adolescents. *Ann. Intern. Med.* **137**, 381–433 (2002).
93. Zhu, H., Bilgin, M. & Snyder, M. Proteomics. *Annu. Rev. Biochem.* **72**, 783–812 (2003).
94. Gluck, R. & Metcalfe, I. C. New technology platforms in the development of vaccines for the future. *Vaccine* **20**, B10–B16 (2002).
95. National Institute of Allergy and Infectious Diseases. *The Jordan Report: 20th Anniversary. Accelerated Development of Vaccines* <http://www.niaid.nih.gov/dmid/vaccines/jordan20/jordan20_2002.pdf> (2002).
96. Bendelac, A. & Medzhitov, R. Adjuvants of immunity: harnessing innate immunity to promote adaptive immunity. *J. Exp. Med.* **195**, F19–F23 (2002).
97. O'Hagan, D. T. & Valiante, N. M. Recent advances in the discovery and delivery of vaccine adjuvants. *Nature Rev. Drug Discov.* **2**, 727–735 (2003).
98. Fraser, C. M., Eisen, J. A., Nelson, K. E., Paulsen, I. T. & Salzberg, S. L. The value of complete microbial genome sequencing (you get what you pay for). *J. Bacteriol.* **184**, 6403–6405 (2002).
99. Cromley, E. K. GIS and disease. *Annu. Rev. Public Health* **24**, 7–24 (2003).

Acknowledgements The authors thank R. M. Krause for helpful discussions, and J. Weddle for graphic design.

Competing interests statement The authors declare that they have no competing financial interests.

Frontal and stealth attack strategies in microbial pathogenesis

D. Scott Merrell¹ & Stanley Falkow²

¹Uniformed Services University of the Health Sciences, Department of Microbiology and Immunology, 4301 Jones Bridge Road, Bethesda, Maryland 20814, USA (e-mail: dmerrell@usuhs.mil)

²Stanford School of Medicine, Department of Microbiology and Immunology, 299 Campus Drive, Fairchild, D051 Stanford, California 94305, USA

Interactions between microbes and human hosts can range from a benign, even symbiotic collaboration to a competition that may turn fatal — resulting in death of the host, the microbe or both. Despite advances that have been made over the past decades in understanding microbial pathogens, more people worldwide still die every year from infectious disease than from any other cause. This highlights the relevance of continuing to probe the mechanisms used by microorganisms to cause disease, and emphasizes the need for new model systems to advance our understanding of host–pathogen interactions.

Although a wide range of microbe–host relationships can ultimately lead to disease, the two most general strategies used by pathogenic microbes may be described in military terms as ‘frontal’ and ‘stealth’ assaults. Pathogenic microorganisms make use of both of these approaches. Typically, frontal assault strategies require that the infecting microbes rapidly replicate, induce disease symptoms that overwhelm the innate defences of the host, and find a new host before engagement of the ‘adaptive’ or ‘acquired’ immune system, in which antigen-specific lymphocytes respond to antigen exposure. Stealth assaults, on the other hand, typically involve a slower infection process in which the microbes subvert the host’s innate and adaptive immune systems to set up a chronic or persistent infection (Box 1). The diverse tactics used for both forms of attack have been the subject of intense investigation, which has helped not only to shape our understanding of the invading organisms, but also to define many of the pathways that are important in host resistance and susceptibility, and in normal host physiology. In discussing microbial pathogenicity, it is important to consider that every human is host to myriad microorganisms, and that illness is the exception rather than the rule (Box 2). Nevertheless, microbial pathogens impose a tremendous medical burden, so it is important that we understand their diverse assault strategies. Here we outline a variety of frontal and stealth strategies used by pathogenic microorganisms. In particular, we discuss the less well-known strategies used by chronically infecting bacteria, and outline the mechanisms used by these pathogens to subvert both the innate and the acquired immune systems.

Frontal assault strategies

Many pathogenic microbes adopt an aggressive strategy in which they immediately attack and attempt to overwhelm the host’s innate immune system. Although there are many variations to this mode of assault (Table 1), these microbes typically produce toxins or use complex secretion systems to deliver so-called effector proteins, which disrupt the normal function of the host cell.

One example of a pathogen that uses a frontal assault strategy is *Vibrio cholerae*, a potent epidemic adversary in many developing areas of the world. The organism is typically ingested in contaminated food and water, and produces

a powerful toxin that causes a voluminous secretory diarrhoea¹. In severely affected patients, the volume of stool purged can be in excess of the patient’s own body weight² and dehydration is the typical cause of death. The ability of *V. cholerae* to rapidly cause disease before eliciting a productive immune response is shared by many microorganisms, including the most common cause of diarrhoeal disease in infants, the rotavirus family, which is responsible for more than 600,000 deaths a year³.

Cholera toxin is encoded in the genome of a bacteriophage, called cholera-toxin phage, that is integrated into the *V. cholerae* genome. Thus, the virulence of *V. cholerae* results from the acquisition, by horizontal genetic transfer, of foreign DNA that encodes a toxin⁴. The receptor used by the cholera-toxin phage to infect *V. cholerae* is encoded as part of an additional portion of horizontally acquired DNA known as a pathogenicity island (PAI).

PAIs, found in many types of bacteria, are large chromosomal regions of DNA that are believed to have originated either from bacteriophages or from small, self-replicating bacterial DNA structures called plasmids⁵. This mechanism, in which acquired DNA has an important role in converting an otherwise harmless species into a pathogen, has been recognized as a common occurrence for many bacteria. It also serves as a microbial example of darwinian natural selection, in which the presence of pathogenic species is partially due to selective pressure applied by the host on the infecting microbe. For example, it is conceivable that before acquisition of the cholera toxin, the ancestral form of *V. cholerae* would colonize and reside in the small intestine but would cause few clinical symptoms. Eventually, this extended stay would result in engagement of the adaptive immune system. Because *V. cholerae* has few, if any, mechanisms to subvert the adaptive branch of the human immune system, most of the bacteria would be killed and not disseminated back to the environment. So although it is possible that the human was a ‘dead end’ for the ancestral form of *V. cholerae*, infection with the cholera-toxin phage produced a strain that could proliferate in the human host and be shed in great numbers back to the aquatic environment before productive engagement of the adaptive immune response. As a result, this pathogen has flourished. This illustrates how host diligence in the form of innate and adaptive immunity imposes selective pressures that shape pathogenic lifestyle, and helps to explain

Box 1

General characteristics of frontal and stealth strategies of pathogenic bacteria

Pathogenic bacteria use numerous strategies to overcome host defences. Broadly speaking, frontal and stealth assault strategies are distinguished by the following characteristics.

Frontal assault

- Short incubation period
- Acute clinical symptoms
- Engages the innate immune system
- Requires transmission to new susceptible hosts to maintain infection
- Rapid microbial replication
- Carrier state is uncommon
- Transmission often requires intimate contact
- Reservoir is often a specific host or the pathogens are opportunistic
- Often induces sterilizing immunity

Stealth assault

- Incubation period may be indeterminate
- Indolent or asymptomatic carriage
- Engages the innate immune system
- Avoids or manipulates the adaptive immune system
- Rapid microbial replication may punctuate a recurrent or terminal event
- Carrier state is common with periods of shedding
- Transmission is by direct contact or persistence in the environment
- Reservoir is usually a specific host or a group of related host species
- Rarely induces sterilizing immunity

why closely related pathogens have such different lifestyles and cause such diverse diseases.

Many bacterial pathogens have been shown to cause disease by delivering proteins directly into the host cell that they are infecting. This is done using specialized type III/IV secretion systems that are typically acquired as part of a PAI. This island encodes components that assemble into a 'molecular syringe', which can inject effector molecules into the host cell^{6,7}. For example, the genus *Yersinia* contains three pathogenic species that use type III secretion to subvert the host immune response and cause diseases ranging from gastrointestinal disorders to bubonic plague. When it enters the host, *Yersinia* encounters macrophages and other specialized phagocytic cells, which serve as the first line of immunological defence. Interaction with these host cells causes *Yersinia* to produce and deliver proteins called *Yersinia* outer-membrane proteins (Yops), which prevent phagocytosis, disrupt normal signal transduction pathways and initiate the ultimate apoptotic demise of the phagocytic cells⁶. Unlike *V. cholerae*, which delivers its potent toxin and is then rapidly shed from the body in the diarrhoeal output, thereby avoiding a strong inflammatory response, *Yersinia* prefers a 'stand and fight' approach in which it faces and overcomes the host's innate and adaptive immune systems.

Stealth assault strategies

Our understanding of the strategies used by pathogens to set up persistent infections remains fairly limited compared with our knowledge of the frontal assault strategies discussed above. It is important to point out that a high proportion of persistent infections are caused by viral and parasitic pathogens^{8–11}. Nevertheless, chronically infecting bacteria present a significant challenge in terms of their ability to cause disease. The remainder of this review will focus primarily on microbes that use stealth assault strategies to cause chronic infections, and on recent findings that may further our understanding of the complex interplay between host and pathogen.

Helicobacter pylori

Infecting over 50% of the world's population, *Helicobacter pylori* has been recognized as something of a bacterial role model for its ability to colonize and persist for the lifetime of the host, even when faced with a robust host immune system. The bacterium exclusively infects human and non-human primates, and persists in the harsh environment of the stomach. About 20% of those infected with *H. pylori* will ultimately suffer diseases ranging from gastritis and ulcer disease to gastric cancer¹². This microbe, which uses a variety of strategies and gene products to subvert the innate and adaptive immune systems of the host (Fig. 1), therefore represents a tremendous medical challenge.

Gastric acidity is a major antimicrobial defence, and the stomach serves as a death chamber for most of the millions of bacteria that are ingested daily. However, *H. pylori* not only survives the low pH, but thrives in the human stomach. The bacterium has a number of strategies to overcome acid stress¹³. Foremost among these approaches is production of the bacterial enzyme urease, which catalyses the hydrolysis of urea to carbon dioxide and ammonia, and helps to maintain a protonmotive force, essential for bacterial metabolism and survival, across the *H. pylori* inner membrane^{14,15}. The production of the basic ammonia molecule helps to buffer the bacterial cytoplasm and the microenvironment as it is secreted from the bacterial cell.

Another important part of the human innate immune system involves Toll-like receptors (TLRs). Bacterial lipopolysaccharide (LPS), a component of the bacterial cell wall, is typically recognized by TLR4 and this triggers a robust proinflammatory response. However, recent work has shown that primary stomach epithelial cells and gastric epithelial cell lines do not react to the LPS from *H. pylori*¹⁶. In addition, it has been shown that recognition of *H. pylori* by TLR5 is markedly different from that observed with other Gram-negative pathogens, as *H. pylori* flagellins do not signal through TLR5 to stimulate an innate immune response^{17,18}. Despite the lack of immunostimulation by these traditional pathways, *H. pylori* is reported to cause a strong inflammatory response *in vivo* through nuclear factor (NF)- κ B activation. It has been suggested that this is required for establishment of a chronic infection¹⁹. So although the bacterium bypasses the innate TLR system, it can engage alternative inflammatory mediators.

H. pylori not only subverts the innate immune system, but also modulates the adaptive immune system by blocking the antigen-dependent proliferation of T cells. This is accomplished partially by delivery of vacuolating cytotoxin, VacA²⁰, which blocks T-cell receptor signalling events that normally lead to the production of cytokines, important mediators of the immune response.

A role in subverting the other branch of the adaptive immune response has been indicated by the finding that ectopic expression in B cells of the CagA protein — a product of the cytotoxin-associated gene (cag) PAI of *H. pylori* — inhibits B-cell proliferation by suppressing the JAK–STAT signalling pathway²¹. This study also suggested that CagA represses B-cell apoptosis, which may contribute to the formation of *H. pylori*-induced mucosa-associated lymphoid tissue (MALT) lymphoma. In this case, host immune defences that are designed to kill the invading microbe are subverted in such a way that they harm the host and not the bacterium.

As well as actively suppressing the immune response and the ability of the host to clear the infection, recent evidence suggests that *H. pylori* can hide from classic host defences by becoming intracellular.

Box 2

Microbial pathogenicity as a survival strategy

Of the 500 or so bacterial species that inhabit us, most never cause illness. But although these 'commensals' (literally 'to eat from the same table') are critical to human existence, some can cause disease in immunologically compromised hosts. This begs the question, what makes some of these flora able to cause disease?

Perhaps the best known of the commensal flora that cause life-threatening disease are those cultured from the human nasopharynx: *Streptococcus pneumoniae*, *Neisseria meningitidis* and *Haemophilus influenzae* type B. Every human is, at one time or another, colonized by these species and most escape unharmed. Although the common assumption is that individuals who become ill do so when their carried strain becomes virulent, in fact the epidemiology of pneumococcal and meningococcal meningitis shows that disease is typically caused by newly acquired strains — even in individuals who had previously been carriers of the same species. The clinical reality is that most people harbour these organisms asymptotically, and disease occurs infrequently in a minority of us.

Descriptions of virulence characteristics of these organisms generally focus on their ability to resist host defence mechanisms by synthesizing antiphagocytic substances, producing specific proteases that degrade host antibodies, and undergoing antigenic variation to avoid recognition by the immune system. However, this overlooks the fact that these 'pathogenicity' determinants probably evolved to facilitate persistent colonization and not disease. But why do certain organisms require these factors to colonize, when the vast majority of our microbial flora do not? One possible explanation is that they have a niche within the nasopharynx that involves crossing an anatomical barrier that keeps other flora at bay. We suggest that this

barrier may be the nasopharyngeal-associated lymphatic tissue (NALT), and that these organisms persist in the small lymph nodes adjacent to the NALT in much the same way as *Salmonella* persists in lymph nodes adjacent to Peyer's patches. This would explain why immunization with the type B capsule of *H. influenzae* protects against both invasive disease and colonization.

This strategy for survival and persistence only pertains to host-adapted microbes. There are also opportunists, which depend on host defence calamity to cause disease. These microbes, usually acquired accidentally from animals, may kill us because we are an evolutionary dead end for them; we have not established the rules of engagement. However, in their host of choice, these same microbes usually behave like our host-adapted flora — with an eye towards persistence rather than towards causing disease and death. The balance shift that causes colonization to go awry, leading to a deadly disease, is a combination of physiological and genetic factors for both participants of the host–pathogen interaction.

In the end, pathogenicity is simply a strategy for microbial survival that involves the inherent ability to breach host barriers that stop other microorganisms. The need to replicate and acquire nutrients provides the driving force for this evolution, which enables the microbe to exploit a previously unoccupied niche. This incursion, and the acquisition of genes that make it possible, has in time become a necessity for survival of these microbes, as they have lost genes that were important for their pre-host-adapted life. The evolution of host–parasite adaptation is still with us, and will continue to be so as long as microorganisms outnumber us numerically and in their replicative power.

Although *H. pylori* is believed to remain primarily extracellular during infection, it was noted by early investigators examining gastric biopsies that a subpopulation of the bacteria could be found within cells^{22,23}. The relevance of this observation was questioned for many years, but recent studies have shown that *H. pylori* can invade cultured eukaryotic cells and reside within multivesicular bodies that promote bacterial survival, motility and replication, and ultimately serve as a reservoir for re-establishment of infection (the intracellular bacteria egress from the vesicles and reseed the extracellular milieu)²⁴. The relevance of these *in vitro* findings is supported by recent work showing that the bacterium could be detected inside precancerous and cancerous epithelial cells from *H. pylori*-positive patients²⁵. In addition, the relative level of expression of several known bacterial virulence factors was considerably higher in the advanced-stage cancerous cells, suggesting that expression of virulence factors by these intracellular bacteria may play a part in disease progression. The ability of *H. pylori* to survive intracellularly and to escape many host immune defences is not unique. It is used by many microbial pathogens, including *Mycobacterium tuberculosis*, *Listeria monocytogenes* and *Shigella flexneri*, the causative agents of tuberculosis, listeriosis and shigellosis, respectively.

A final mechanism that may be crucial for *H. pylori* to escape the host immune response involves its ability to undergo genetic rearrangement that either eliminates particular immunostimulatory gene products or causes variation in potentially immunostimulatory molecules. Genotyping of clonal isolates using an *H. pylori* microarray from a single human host suggests that as many as 3% of the loci show significant genetic variation²⁶. Although the implications of this finding are not fully understood, a study showing that strains of *H. pylori* that have deleted all or portions of the PAI that delivers CagA to the host cell are more likely to colonize various animal models suggests that the ability of the bacterium to modulate gene products that affect the robustness of the immune response is significant²⁷.

Although CagA is a virulence factor for *H. pylori*, the presence of CagA is not sufficient to predict the disease outcome in *H. pylori* infection. It is known that individuals infected with CagA that can be phosphorylated have more severe disease than patients infected with a non-phosphorylatable CagA, but which of these forms of CagA was the progenitor in the evolution of *H. pylori* is unclear. A functional PAI is found in most human clinical isolates, but there is variation in the *cagA* coding region in the critical EPIYA domains that are phosphorylated by host tyrosine kinases²⁸.

These epidemiological results are supported by the recent finding that intragenomic recombination in the *cagA* coding sequence results in the production of a non-phosphorylatable form of CagA that does not induce morphological changes associated with pathogenicity in host cells²⁹. The importance of this strategy of genetic rearrangement within the host may explain the recent identification of the RuvC protein, a Holliday junction resolvase involved in recombination, as a crucial factor in bacterial persistence³⁰.

Many other pathogens use genetic rearrangement as a means of immune escape. Included among these are *Neisseria gonorrhoeae* and HIV, which are responsible for gonorrhoea and AIDS, respectively.

Salmonella enterica serovar Typhi

Many *Salmonella* serovars, like *Salmonella typhi*, the causative agent of typhoid fever, are adapted exclusively to human hosts and are endemic in areas of the world with poor sanitation and inadequate sewage-treatment facilities. Typhoid fever remains an important cause of human morbidity and mortality, and is becoming increasingly difficult to treat owing to the emergence of antibiotic-resistant strains³¹. Between 1% and 6% of those infected with *S. typhi* become chronic carriers of the bacterium and, although not necessarily suffering any symptoms themselves, may serve as a reservoir for subsequent infections during periods of faecal and urinal shedding³² — think Typhoid Mary (see Box 3). Sites of persistence within the body have been shown to include the bone marrow and the gall bladder^{33,34}, but little

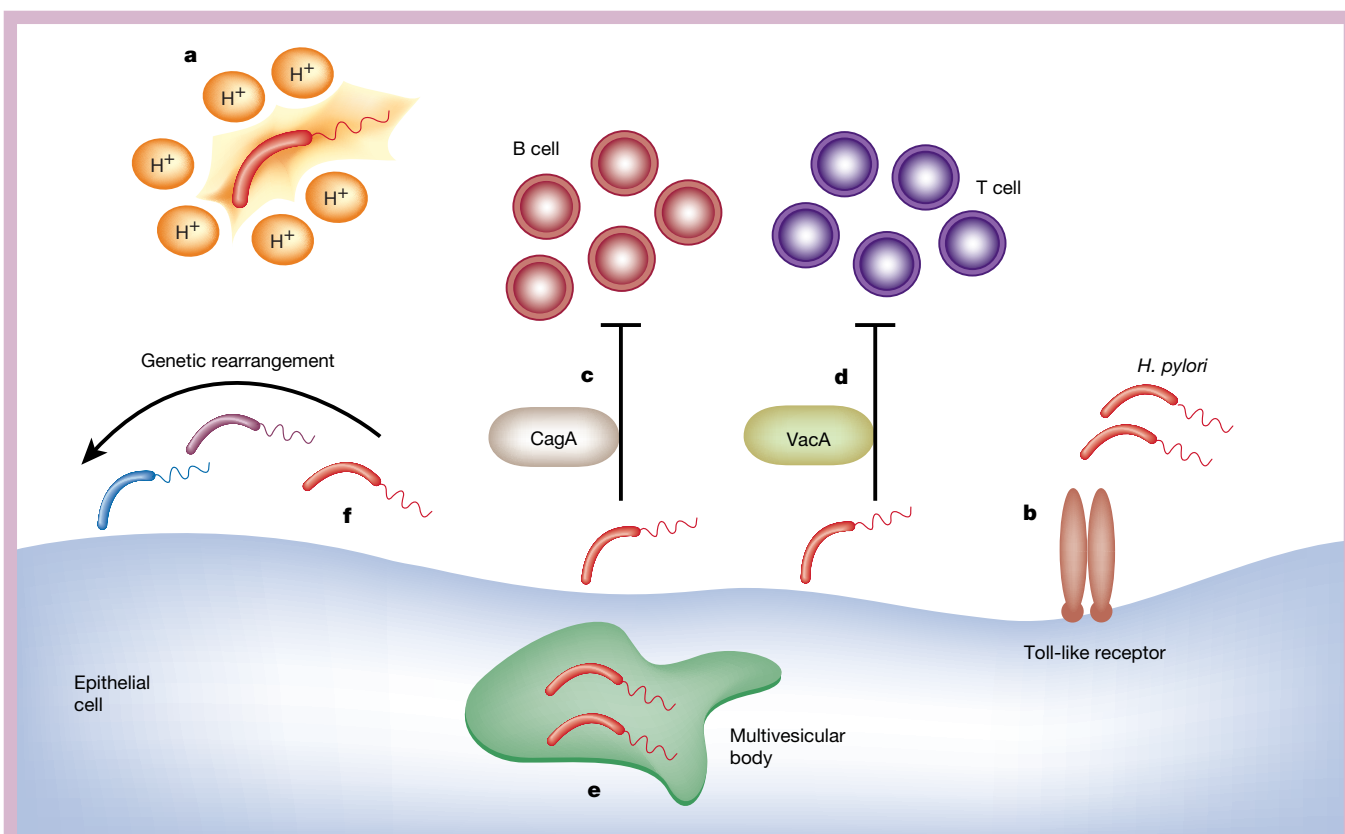


Figure 1 Mechanisms used by *Helicobacter pylori* to establish persistent infection. The bacterium overcomes the innate immune system by neutralizing gastric acid (a) and by having lipopolysaccharide and flagella with low immunostimulatory recognition by Toll-like receptors (b). The adaptive immune system is also affected: B-cell (c) and T-cell (d) proliferation are blocked by the activity of CagA (a product of the cytotoxin-associated gene (*cag*) pathogenicity island) and VacA (vacuolating cytotoxin), respectively. In addition, the bacterium can escape exposure to host immune defences by surviving intracellularly (e), and shows a high frequency of genetic rearrangement (f) that appears to be essential for persistent colonization.

is understood about the adaptive mechanisms used by the bacteria during the course of establishing a persistent infection. Recent work using a novel model of bacterial persistence with *S. typhimurium* has begun to shed light on *Salmonella* persistence strategies³⁵ and should be amenable to genetic analysis, both to dissect the role of individual bacterial factors in persistence and to explain the failure of the host immune system to eradicate the infection.

Adaptation of *Salmonella* serovars to the host and the capacity of adapted microbes to persist are essential survival features of most, if not all, *Salmonella* serovars. Although *S. typhimurium* is well known as a cause of food-borne gastroenteritis in humans, this contributes little to the survival of the species because it rarely establishes a carrier state in humans and is associated only occasionally with intrafamilial spread. Mice infected with *S. typhimurium* have been used as a model to study acute systemic disease, as the bacterium causes a typhoid-like disease in some mouse strains. Differences in mouse susceptibility to *S. typhimurium* have been linked to the particular allele of the *Nramp1* gene being expressed in cells of the monocyte/macrophage lineage³⁶. *Nramp1* is involved in controlling exponential growth of the bacteria in the reticuloendothelial system during the early stages of infection. In a recent study, mice expressing the wild-type *Nramp1* allele did not die after oral inoculation with *S. typhimurium*, but became uniformly persistently infected in a manner similar to that observed in chronic typhoid carriers³⁵. In these animals, *Salmonella* was found to persist predominately in the mesenteric lymph nodes and/or spleen even in the face of a robust antibody response. Closer inspection showed that the bacteria were actually intracellular, having colonized macrophages. Most of the infected macrophages only contained between one and three bacteria — fewer than would be expected on the basis of studies in cultured macrophages, in which

the SPI-2 pathogenicity island of *Salmonella* (a genetic element responsible for causing disease) is thought to permit intracellular replication. Whether SPI-2 is actually expressed in persistent infection of macrophages is not certain. However, reactivation of intracellular *Salmonella* and systemic spread could be accomplished by administering interferon- γ (IFN- γ)-neutralizing antibodies, suggesting that the host cytokine IFN- γ is important for suppression of *Salmonella* replication and disease³⁵. IFN- γ is also important for the maintenance of latent infections of *Mycobacterium tuberculosis* and *Chlamydia trachomatis*. Using this persistence model and recent technological advances, it should be possible to identify bacterial factors that are important for persistence. This could be accomplished using microarrays and global screening strategies of libraries of defined *Salmonella* mutants³⁷.

Bartonella spp.

The *Bartonella* genus is increasingly recognized as a significant human pathogen. *Bartonella* spp. cause numerous diseases, including Oroya fever, verruga peruana, trench fever, bacillary angiomatosis, endocarditis and cat-scratch disease³⁸. *Bartonella* infections are disseminated from their natural reservoir (cats, rats, humans, deer and other mammals) by arthropods such as fleas and ticks. Carrier rates in the reservoir population are quite high: as many as 41% of domestic cats have been shown to be infected with *B. henselae*³⁹. *Bartonellae* are fairly unique among bacterial pathogens in terms of their ability to sustain prolonged periods of parasitism within red blood cells. They have been shown to invade, multiply within and persist for the lifetime of the infected host cell⁴⁰, and to reach titres of 10⁴ per ml of blood in infected humans⁴¹. How this many Gram-negative organisms can persist within the bloodstream without inducing a classic

Table 1 An overview of mechanisms used by bacteria to avoid host immune responses

Immune response affected	Pathogen	Bacterial factor/mechanism	References
Induction of apoptosis/cytotoxicity	<i>Shigella flexneri</i> <i>Yersinia</i>	IpaB activation of pro-caspase 1 YopJ/P injection into macrophages	50 6
Inhibition of apoptosis	<i>Chlamydia</i> <i>Mycobacterium tuberculosis</i>	IL-10-induced TNF- α suppression Increased expression of <i>bcl2</i> and <i>Rb</i>	51 52
Inhibition of cytokine production	<i>Bartonella</i> <i>Vibrio cholerae</i>	IL-10 induction suppresses other cytokines CT inhibition of IL-12 secretion	43 53
Cytokine overproduction	<i>Bordetella pertussis</i> <i>Helicobacter pylori</i>	Pertussis toxin induction of IL-1 and IL-4 IL-8 induction	54, 55 12
Complement evasion	<i>Streptococcus pyogenes</i> <i>Porphyromonas gingivalis</i>	M protein binding of C4BP Protease inactivation of C3 and C5	56 57
Evasion of host antibodies	<i>Staphylococcus aureus</i> <i>Peptostreptococcus magnus</i>	Protein A binding of IgG blocks phagocytosis Protein L binding of κ light chains	58 59
Inhibition of antigen presentation	<i>Helicobacter pylori</i> <i>Mycobacterium tuberculosis</i>	VacA targeting of antigen-presenting cells Downregulation of MHC class II and CD1	20 60
Inhibition of phagocytosis	<i>Pseudomonas</i> <i>Yersinia</i>	ExoT, ExoS targeting of Rho, Rac and Cdc42 YopO targeting of RhoA, Rac and actin	61, 62 6
Survival in phagocytes	<i>Legionella</i> <i>Coxiella burnetii</i>	Dot/ICM genes inhibit phagolysosome fusion Acid tolerance allows survival in acidified lysosomes	63, 64 65
Phase/antigenic variation	<i>Neisseria gonorrhoeae</i> <i>Helicobacter pylori</i>	Modulation of pilin Modulation of PAI and CagA	66 27, 29

CagA, product of cytotoxin-associated gene (cag) PAI; CT, cholera toxin; ICM, intracellular multiplication; IgG, immunoglobulin- γ ; IL, interleukin; MHC, major histocompatibility complex; PAI, pathogenicity island; TNF- α , tumour-necrosis factor- α ; VacA, vacuolating cytotoxin.

septic shock response remains to be determined, but the finding suggests that, as with *H. pylori*, the LPS of *Bartonella* may have reduced immunostimulatory properties. Also, as for *Salmonella*, the ability to survive intracellularly helps *Bartonella* to escape the host immune response. Antibodies have no effect on *Bartonella* within red blood cells, but may contribute to preventing new waves of bacterial invasion⁴². It has also been suggested that intracellular *Bartonella* may affect the development of an adaptive immune response in a manner similar to that seen with malaria-infected red blood cells⁴³. Malaria prevents the maturation of dendritic cells, which are essential for the development of a normal adaptive immune response⁴⁴.

Genetic studies and development of appropriate animal models of haematropic infection by *Bartonella* have begun to shed light on some of the bacterial factors that are required for intraerythrocytic infection (Fig. 2). One important virulence factor in *Bartonella* is the type IV secretion system. It was recently shown that mutation of either *virB4* or *virD4* genes in *B. tribocorum* prevents the bacterium from establishing intraerythrocytic bacteraemia in a rat model of infection⁴⁵ — although whether this failure is due to changes in endothelial cell interaction or the inability to deliver a necessary effector protein remains to be determined. It is notable that another family of zoonotically acquired human pathogens, *Brucella*, which causes undulant fever in humans, also requires a type IV secretion system to establish infection⁴⁶.

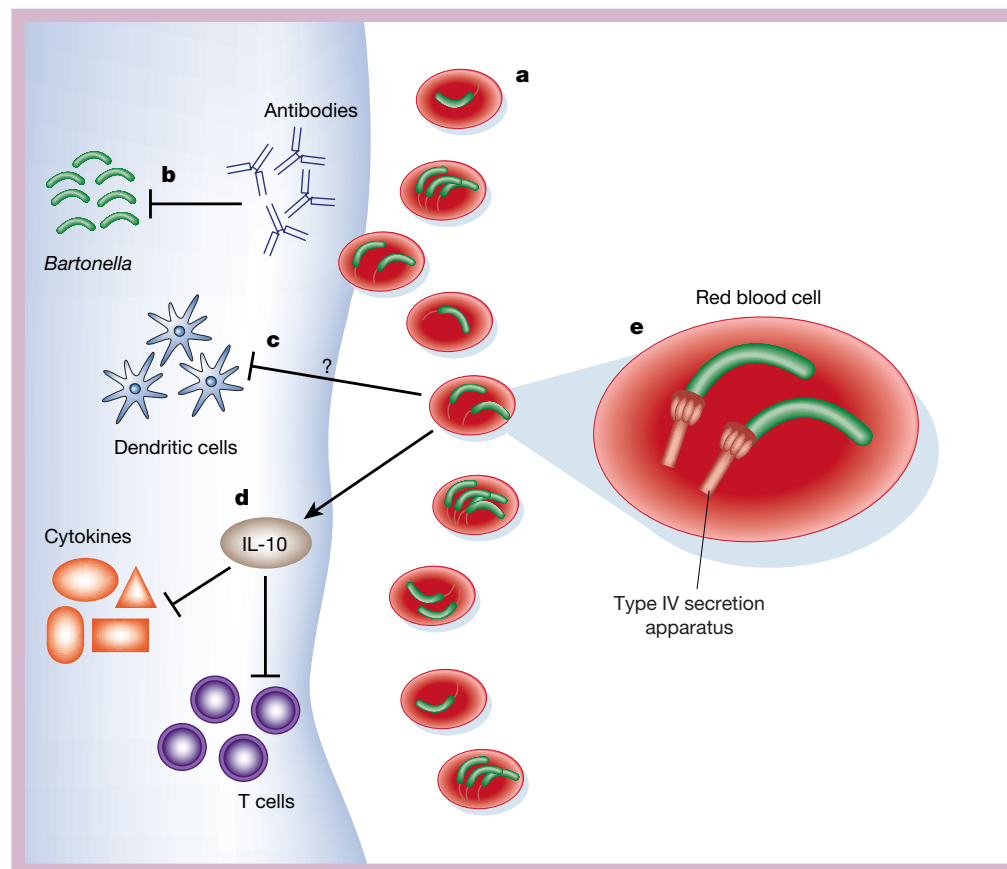


Figure 2 Proposed strategies by which *Bartonella* establishes persistent infection. **a**, *Bartonella* colonizes an unknown niche within the body and seeds bacteria that are able to infect and survive inside red blood cells. **b**, Host antibodies have no effect on intraerythrocytic bacteria but prevent new waves of erythrocytic parasitism. **c, d**, Intracellular bacteria have been proposed to affect dendritic cell maturation through an unknown mechanism (**c**), and have been shown to inhibit cytokine production and T-cell proliferation (**d**) by inducing the expression of interleukin-10 (IL-10). **e**, A type IV secretion system is essential for establishing persistent intraerythrocytic infection.

Box 3

What's in a name?

It's unlikely that you've ever heard the name Mary Mallon. Mary was a young Irish woman who emigrated to the United States in 1883 in search of a better life. You may be familiar with 'Typhoid Mary', the moniker by which Mary Mallon came to be known. The legend of Typhoid Mary has grown over the years and she is reputed to have caused hundreds, if not thousands, of deaths in turn-of-the-century New York. In reality, she was probably responsible for only 33 cases of typhoid fever, resulting in three deaths.

Typhoid fever is caused by the bacterium *Salmonella typhi* and is characterized by a high fever, powerful headaches, nausea, diarrhoea and tenderness of the abdomen. The bacterium can be spread easily between humans via a faecal-oral route, and in the 1800s typhoid killed one in every ten people who were infected and so was feared perhaps as much as the plague.

However, *S. typhi* also has the ability to colonize some people without causing any disease symptoms. Unknown to Mary, who worked as a cook for a number of different families, she was a carrier of this deadly bacterium. Because *S. typhi* can be spread easily,

Mary's constant handling of uncooked foods with improperly cleaned hands led her to cause typhoid outbreaks in seven of the eight families that she worked for, as well as in a hospital where she worked for a short while. Her carrier status was discovered in 1906 by George Soper, a sanitary engineer who was hired to determine the source of the mysterious typhoid outbreaks that seemed to appear wherever Mary was working.

As one of the first recorded typhoid carriers in the United States, Mary became reviled for her inability to understand how she could be spreading disease when she wasn't unwell herself, and for her refusal to stop cooking, which was her only means of financial support at the time. Eventually, she was exiled to the Riverside Hospital on North Brother Island, New York, where she lived in virtual isolation for 26 years and died of complications from a stroke in 1938. Her memory has lived on, however, and now the name Typhoid Mary is synonymous with death and disease, despite the fact that Mary Mallon started out as a simple, apparently healthy woman who was just trying to make a living.

In addition to its ability to invade and persist within red blood cells, *Bartonella* apparently manipulates the host immune system through modification of host-cell cytokine production. It was shown that homeless people who were chronically and asymptotically infected with *B. quintana* showed decreased levels of circulating markers of leukocyte activation and decreased cytokine production by mononuclear cells. This was accompanied by an increase in the secretion of interleukin-10 (IL-10)⁴³, which depresses the expression of other cytokines and the release of soluble inflammatory mediators. Activation of T cells in the presence of IL-10 has been shown to result in non-responsiveness of these cells and attenuation of downstream signalling events⁴⁷. Thus *Bartonella*-induced expression of IL-10 may directly facilitate bacterial persistence. This increasingly recognized persistence strategy has been observed for *Leishmania major*, *Coxiella burnetii*, *Yersinia pseudotuberculosis* and *Onchocerca volvulus*, the causative agents of leishmaniasis, Q fever, gastroenteritis and river blindness, respectively. Like *Bartonella*, these pathogens not only circumvent normal host defences, but also directly manipulate them.

Fighting back

Years of research have helped to define our current understanding of pathogens and host-pathogen interactions. But of the more than 1,400 known bacterial, viral and parasitic pathogens that infect humans, we have managed to completely eradicate only smallpox⁴⁸. Although the spread of several pathogens has been greatly diminished over the past 30 years, a further 37 human pathogens have been discovered and an estimated 12% of the known pathogens have been recognized as emerging or re-emerging⁴⁹. About 200 different bacterial species are known to cause human disease⁴⁸, and these microbes use multiple and diverse strategies to overcome the host immune system. With the continuing rise in infectious-disease-related morbidity and mortality, future research should be directed towards gaining a greater understanding of the host-pathogen interface and strategies used by microbes for modulation of the host immune system. To accomplish this, we must focus on clarifying not only virulence factors *per se*, but also specific mechanisms that allow microbes to persist within the host environment. Gaining a thorough knowledge of chronically infecting microbes and learning how to control them is perhaps the greatest challenge. It is our belief that this should be made a major research focus and funding priority. We must strive to develop and apply novel technological approaches and new model systems to the elucidation of persistence strategies if we ever hope to find ways to subvert these stealthy pathogens. □

doi:10.1038/nature02760

- Kaper, J. B., Fasano, A. & Trucksis, M. in *Vibrio cholerae and Cholera: Molecular to Global Perspectives* (eds Wachsmuth, I. K., Blake, P. A. & Olsvik, Ø.) 145–176 (ASM Press, Washington DC, 1994).
- Phillips, R. A. Water and electrolyte losses in cholera. *Fed. Proc.* **23**, 705–718 (1964).
- Wilhelmi, I., Roman, E. & Sanchez-Fauquier, A. Viruses causing gastroenteritis. *Clin. Microbiol. Infect.* **9**, 247–262 (2003).
- Waldor, M. K. & Mekalanos, J. J. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* **272**, 1910–1914 (1996).
- Schmidt, H. & Hensel, M. Pathogenicity islands in bacterial pathogenesis. *Clin. Microbiol. Rev.* **17**, 14–56 (2004).
- Cornelis, G. R. The *Yersinia* Ysc–Yop 'type III' weaponry. *Nature Rev. Mol. Cell Biol.* **3**, 742–752 (2002).
- Christie, P. J. & Vogel, J. P. Bacterial type IV secretion: conjugation systems adapted to deliver effector molecules to host cells. *Trends Microbiol.* **8**, 354–360 (2000).
- Brake, D. A. Parasites and immune responses: memory illusion? *DNA Cell Biol.* **22**, 405–419 (2003).
- Crowe, S., Zhu, T. & Muller, W. A. The contribution of monocyte infection and trafficking to viral persistence, and maintenance of the viral reservoir in HIV infection. *J. Leukoc. Biol.* **74**, 635–641 (2003).
- Jarvis, M. A. & Nelson, J. A. Human cytomegalovirus persistence and latency in endothelial cells and macrophages. *Curr. Opin. Microbiol.* **5**, 403–407 (2002).
- Johnson, W. E. & Desrosiers, R. C. Viral persistence: HIV's strategies of immune system evasion. *Annu. Rev. Med.* **53**, 499–518 (2002).
- Ernst, P. B. & Gold, B. D. The disease spectrum of *Helicobacter pylori*: the immunopathogenesis of gastroduodenal ulcer and gastric cancer. *Annu. Rev. Microbiol.* **54**, 615–640 (2000).
- Valenzuela, M., Cerda, O. & Toledo, H. Overview on chemotaxis and acid resistance in *Helicobacter pylori*. *Biol. Rev.* **36**, 429–436 (2003).
- Weeks, D. L., Eskandari, S., Scott, D. R. & Sachs, G. A H⁺-gated urea channel: the link between *Helicobacter pylori* urease and gastric colonization. *Science* **287**, 482–485 (2000).
- Mobley, H. L., Island, M. D. & Hausinger, R. P. Molecular biology of microbial ureases. *Microbiol. Rev.* **59**, 451–480 (1995).
- Smith, M. F. Jr *et al.* Toll-like receptor (TLR) 2 and TLR5, but not TLR4, are required for *Helicobacter pylori*-induced NF- κ B activation and chemokine expression by epithelial cells. *J. Biol. Chem.* **278**, 32552–32560 (2003).
- Lee, S. K. *et al.* *Helicobacter pylori* flagellins have very low intrinsic activity to stimulate human gastric epithelial cells via TLR5. *Microbes Infect.* **5**, 1345–1356 (2003).
- Gewirtz, A. T. *et al.* *Helicobacter pylori* flagellin evades Toll-like receptor 5-mediated innate immunity. *J. Infect. Dis.* **189**, 1914–1920 (2004).
- Rhen, M., Eriksson, S., Clements, M., Bergstrom, S. & Normark, S. J. The basis of persistent bacterial infections. *Trends Microbiol.* **11**, 80–86 (2003).
- Gebert, B., Fischer, W., Weiss, E., Hoffmann, R. & Haas, R. *Helicobacter pylori* vacuolating cytotoxin inhibits T lymphocyte activation. *Science* **301**, 1099–1102 (2003).
- Umehara, S., Higashi, H., Ohnishi, N., Asaka, M. & Hatakeyama, M. Effects of *Helicobacter pylori* CagA protein on the growth and survival of B lymphocytes, the origin of MALT lymphoma. *Oncogene* **22**, 8337–8842 (2003).
- Bode, G., Malfertheiner, P. & Ditschuneit, H. Pathogenetic implications of ultrastructural findings in *Campylobacter pylori* related gastroduodenal disease. *Scand. J. Gastroenterol. Suppl.* **142**, 25–39 (1988).
- Foliguet, B., Vicari, E., Guedenet, J. C., De Korwin, J. D. & Marchal, L. Scanning electron microscopic study of *Campylobacter pylori* and associated gastroduodenal lesions [in French]. *Gastroenterol. Clin. Biol.* **13**, 65B–70B (1989).
- Amieva, M. R., Salama, N. R., Tompkins, L. S. & Falkow, S. *Helicobacter pylori* enter and survive within multivesicular vacuoles of epithelial cells. *Cell. Microbiol.* **4**, 677–690 (2002).
- Semino-Mora, C. *et al.* Intracellular and interstitial expression of *Helicobacter pylori* virulence genes in gastric precancerous intestinal metaplasia and adenocarcinoma. *J. Infect. Dis.* **187**, 1165–1177 (2003).
- Israel, D. A. *et al.* *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. *Proc. Natl. Acad. Sci. USA* **98**, 14625–14630 (2001).

27. Philpott, D. J. *et al.* Reduced activation of inflammatory responses in host cells by mouse-adapted *Helicobacter pylori* isolates. *Cell. Microbiol.* **4**, 285–296 (2002).
28. Azuma, T. *et al.* Association between diversity in the Src homology 2 domain-containing tyrosine phosphatase binding site of *Helicobacter pylori* CagA protein and gastric atrophy and cancer. *J. Infect. Dis.* **189**, 820–827 (2004).
29. Aras, R. A. *et al.* Natural variation in populations of persistently colonizing bacteria affect human host cell phenotype. *J. Infect. Dis.* **188**, 486–496 (2003).
30. Loughlin, M. F., Barnard, F. M., Jenkins, D., Sharples, G. J. & Jenks, P. J. *Helicobacter pylori* mutants defective in RuvC Holliday junction resolvase display reduced macrophage survival and spontaneous clearance from the murine gastric mucosa. *Infect. Immun.* **71**, 2022–2031 (2003).
31. Wain, J. *et al.* Molecular typing of multiple-antibiotic-resistant *Salmonella enterica* serovar Typhi from Vietnam: application to acute and relapse cases of typhoid fever. *J. Clin. Microbiol.* **37**, 2466–2472 (1999).
32. Levine, M. M., Black, R. E. & Lanata, C. Precise estimation of the numbers of chronic carriers of *Salmonella typhi* in Santiago, Chile, an endemic area. *J. Infect. Dis.* **146**, 724–726 (1982).
33. Wain, J. *et al.* Quantitation of bacteria in bone marrow from patients with typhoid fever: relationship between counts and clinical features. *J. Clin. Microbiol.* **39**, 1571–1576 (2001).
34. Sinnott, C. R. & Teall, A. J. Persistent gallbladder carriage of *Salmonella typhi*. *Lancet* **1** (8539), 976 (1987).
35. Monack, D. M., Bouley, D. M. & Falkow, S. *Salmonella typhimurium* persists within macrophages in the mesenteric lymph nodes of chronically infected *Nramp1*^{−/−} mice and can be reactivated by IFN- γ neutralization. *J. Exp. Med.* **199**, 231–241 (2004).
36. Vidal, S. M., Malo, D., Vogan, K., Skamene, E. & Gros, P. Natural resistance to infection with intracellular parasites: isolation of a candidate for Bcg. *Cell* **73**, 469–485 (1993).
37. Merrell, D. S. & Camilli, A. Information overload: assigning genetic functionality in the age of genomics and large-scale screening. *Trends Microbiol.* **10**, 571–574 (2002).
38. Koehler, J. E. in *Persistent Bacterial Infections* (eds Nataro, J. M. & Blaser, M. J.) 339–353 (ASM Press, Washington DC, 2000).
39. Koehler, J. E., Glaser, C. A. & Tappero, J. W. *Rochalimaea henselae* infection. A new zoonosis with the domestic cat as reservoir. *J. Am. Med. Assoc.* **271**, 531–535 (1994).
40. Schulein, R. *et al.* Invasion and persistent intracellular colonization of erythrocytes. A unique parasitic strategy of the emerging pathogen *Bartonella*. *J. Exp. Med.* **193**, 1077–1086 (2001).
41. Koehler, J. E., Quinn, F. D., Berger, T. G., LeBoit, P. E. & Tappero, J. W. Isolation of *Rochalimaea* species from cutaneous and osseous lesions of bacillary angiomatosis. *N. Engl. J. Med.* **327**, 1625–1631 (1992).
42. Koesling, J., Aebischer, T., Falch, C., Schulein, R. & Dehio, C. Cutting edge: antibody-mediated cessation of hemotropic infection by the intraerythrocytic mouse pathogen *Bartonella grahamii*. *J. Immunol.* **167**, 11–14 (2001).
43. Capo, C., Amirayan-Chevillard, N., Brouqui, P., Raoult, D. & Mege, J. L. *Bartonella quintana* bacteremia and overproduction of interleukin-10: model of bacterial persistence in homeless people. *J. Infect. Dis.* **187**, 837–844 (2003).
44. Urban, B. C. *et al.* *Plasmodium falciparum*-infected erythrocytes modulate the maturation of dendritic cells. *Nature* **400**, 73–77 (1999).
45. Schulein, R. & Dehio, C. The VirB/VirD4 type IV secretion system of *Bartonella* is essential for establishing intraerythrocytic infection. *Mol. Microbiol.* **46**, 1053–1067 (2002).
46. Watarai, M., Makino, S. & Shirahata, T. An essential virulence protein of *Brucella abortus*, VirB4, requires an intact nucleoside-triphosphate-binding domain. *Microbiology* **148**, 1439–1446 (2002).
47. Jenkins, M. K. *et al.* *In vivo* activation of antigen-specific CD4 T cells. *Annu. Rev. Immunol.* **19**, 23–45 (2001).
48. Mascie-Taylor, C. G. & Karim, E. The burden of chronic disease. *Science* **302**, 1921–1922 (2003).
49. Gubler, D. J. Resurgent vector-borne diseases as a global health problem. *Emerg. Infect. Dis.* **4**, 442–450 (1998).
50. Hilbi, H. *et al.* Shigella-induced apoptosis is dependent on caspase-1 which binds to IpaB. *J. Biol. Chem.* **273**, 32895–32900 (1998).
51. Geng, Y. *et al.* *Chlamydia pneumoniae* inhibits apoptosis in human peripheral blood mononuclear cells through induction of IL-10. *J. Immunol.* **164**, 5522–5529 (2000).
52. Danelishvili, L., McGarvey, J., Li, Y. J. & Bermudez, L. E. *Mycobacterium tuberculosis* infection causes different levels of apoptosis and necrosis in human macrophages and alveolar epithelial cells. *Cell. Microbiol.* **5**, 649–660 (2003).
53. Braun, M. C., He, J., Wu, C. Y. & Kelsall, B. L. Cholera toxin suppresses interleukin (IL)-12 production and IL-12 receptor β 1 and β 2 chain expression. *J. Exp. Med.* **189**, 541–552 (1999).
54. Vistica, B. P., McAllister, C. G., Sekura, R. D., Ihle, J. N. & Gery, I. Dual effects of pertussis toxin on lymphoid cells in culture. *Cell. Immunol.* **101**, 232–241 (1986).
55. Mu, H. H. & Sewell, W. A. Enhancement of interleukin-4 production by pertussis toxin. *Infect. Immun.* **61**, 2834–2840 (1993).
56. Thern, A., Stenberg, L., Dahlback, B. & Lindahl, G. Ig-binding surface proteins of *Streptococcus pyogenes* also bind human C4b-binding protein (C4BP), a regulatory component of the complement system. *J. Immunol.* **154**, 375–386 (1995).
57. Grenier, D. Inactivation of human serum bactericidal activity by a trypsinlike protease isolated from *Porphyromonas gingivalis*. *Infect. Immun.* **60**, 1854–1857 (1992).
58. Peterson, P. K., Verhoef, J., Sabath, L. D. & Quie, P. G. Effect of protein A on staphylococcal opsonization. *Infect. Immun.* **15**, 760–764 (1977).
59. Bjorck, L. Protein L. A novel bacterial cell wall protein with affinity for Ig L chains. *J. Immunol.* **140**, 1194–1197 (1988).
60. Stenger, S., Niaz, K. R. & Modlin, R. L. Down-regulation of CD1 on antigen-presenting cells by infection with *Mycobacterium tuberculosis*. *J. Immunol.* **161**, 3582–3588 (1998).
61. Krall, R., Schmidt, G., Aktories, K. & Barbieri, J. T. *Pseudomonas aeruginosa* ExoT is a Rho GTPase-activating protein. *Infect. Immun.* **68**, 6066–6068 (2000).
62. Barbieri, J. T. *Pseudomonas aeruginosa* exoenzyme S, a bifunctional type-III secreted cytotoxin. *Int. J. Med. Microbiol.* **290**, 381–387 (2000).
63. Marra, A., Blander, S. J., Horwitz, M. A. & Shuman, H. A. Identification of a *Legionella pneumophila* locus required for intracellular multiplication in human macrophages. *Proc. Natl Acad. Sci. USA* **89**, 9607–9611 (1992).
64. Berger, K. H. & Isberg, R. R. Two distinct defects in intracellular growth complemented by a single genetic locus in *Legionella pneumophila*. *Mol. Microbiol.* **7**, 7–19 (1993).
65. Baca, O. G., Li, Y. P. & Kumar, H. Survival of the Q fever agent *Coxiella burnetii* in the phagolysosome. *Trends Microbiol.* **2**, 476–480 (1994).
66. Gray-Owen, S. D., Dehio, C., Rudel, T., Naumann, M. & Meyer, T. F. in *Principles of Bacterial Pathogenesis* (ed. Groisman, E. A.) 559–618 (Academic Press, San Diego, 2001).

Acknowledgements The authors thank J. Koehler, L. Thompson, A. Mueller and A. Camilli for comments on the manuscript. Research in the laboratories of D.S.M. and S.F. are supported by funds from the Damon Runyon Foundation and the National Institutes of Health.

Competing interests statement The authors declare that they have no competing financial interests.

Inferences, questions and possibilities in Toll-like receptor signalling

Bruce Beutler

Department of Immunology, The Scripps Research Institute, 10550 N Torrey Pines Road, La Jolla, California 92037, USA
(e-mail: bruce@scripps.edu)

The Toll-like receptors (TLRs) are the key proteins that allow mammals — whether immunologically naive or experienced — to detect microbes. They lie at the core of our inherited resistance to disease, initiating most of the phenomena that occur in the course of infection. Quasi-infectious stimuli that have been used for decades to study inflammatory mechanisms can activate the TLR family of proteins. And it now seems that many inflammatory processes, both sterile and infectious, may depend on TLR signalling. We are in a good position to apply our understanding of TLR signalling to a range of challenges in immunology and medicine.

Two key discoveries guided the field of innate immunity to its present era of accelerated advancement. In 1996, the *Drosophila* protein Toll, formerly known only for its developmental functions¹, was shown to be required for flies to mount an effective immune response to the fungus *Aspergillus fumigatus*². In 1998, TLR4 was positionally identified as the lipopolysaccharide (LPS) receptor, encoded by the *Lps* locus and known to be required for mice to mount effective responses to Gram-negative bacteria, in which LPS is an integral part of the outer cell membrane³. These findings revealed that the proximal innate immune sensing apparatus of insects and mammals are related by descent, and suggested a central role for TLRs in the primary recognition of infectious pathogens by mammals.

Thirteen mammalian TLR paralogues have now been identified^{4,5} (10 in humans and 12 in mice). An early hypothesis held that each TLR senses a distinct repertoire of conserved microbial molecules, so that collectively, the TLRs can detect most microbes, and perhaps all of them. Both 'forward' and 'reverse' genetic tools have been applied to determine the molecular specificity of most of the TLRs⁶, and to an amazing extent, the hypothesis just stated has been validated. There is no longer any doubt that the TLRs are capable of sensing organisms ranging from protozoa^{7,8} to bacteria⁹⁻¹² to fungi¹³ to viruses^{4,14-18}.

This, in brief, is why the TLRs are so important, and why the TLR field has been so explosive. The TLRs have filled a void in immunology. The impact of their identification as the gateway to innate immune recognition can be compared to earlier discoveries of the receptors that serve 'adaptive' or 'acquired' immune recognition — the immunoglobulins and the T-cell receptors. But to the extent that innate immune mechanisms precede and empower the adaptive immune response, the TLRs are perhaps the most fundamental and essential immune receptors of all. What, then, can we do with our new-found understanding?

Here I review the state of play in the field of TLR signalling, concluding that advances in our understanding of TLR pathways may allow us to tackle a range of challenges in immunology and medicine, including autoimmunity and sepsis.

The new questions

In large part, the specificities of the TLRs have been deciphered, and a good deal is known about how they signal. Four adaptor proteins — MyD88 (myeloid differentiation

factor 88), MAL/TIRAP (MyD88-adaptor-like/TIR-associated protein), TRIF (Toll-receptor-associated activator of interferon) and TRAM (Toll-receptor-associated molecule) — transduce signals from all of the TIR (Toll/interleukin-1 receptor (IL-1R) homologous region) domains, activating protein kinases and then transcription factors that cause inflammatory effects (Fig. 1). The function of a fifth adaptor, SARM (sterile alpha and HEAT/Armadillo motif protein), has yet to be defined. At the margins of what we now understand, some entirely new questions are being asked, and some of the earlier questions have been turned upside down. We are no longer entirely ignorant about the nature of innate immune receptors, nor about how we detect pathogens. Instead, where infectious disease is concerned, are there any organisms that the TLRs do not sense? Maybe helminths; maybe certain viruses; and maybe a number of other organisms, although such negative conclusions cannot be drawn with certainty. Concomitantly, are there primary mechanisms for sensing microbes that do not involve TLRs or elements of their signalling pathways? Perhaps the NOD (nucleotide-binding oligomerization domain) proteins¹⁹; certainly some of the receptors on natural killer cells²⁰; and where double-stranded RNA is concerned, other proteins as well²¹. Complement and mannose-binding proteins are also selective in their ability to recognize microbes. But already, there are clear examples of interplay between TLR-mediated signalling and these alternative pathways. For example, an effective response to mouse cytomegalovirus requires the protein Ly49H as well as TLR3 and TLR9 (ref. 4).

In the area of adaptive immune responses, do the TLRs (and the dendritic cells that use them for sensing) dictate the type of adaptive response that will occur when an antigen is presented? It has been proposed that TLRs initiate responses involving helper T cells of the T_H1 variety, whereas an entirely different class of receptors, so far elusive, initiates T_H2 responses²²; T_H1 and T_H2 cells produce different groups of cytokines and support different immune functions. Perhaps most interesting of all, beyond the realm of infectious diseases *per se*, is the question, do the TLRs and their signalling pathways contribute to autoimmunity and/or to sterile inflammation?

The old challenges

All of the questions just posed are related to one another by virtue of the fact that quasi-infectious stimuli (for example, LPS or heat-killed microbes) have been used for decades

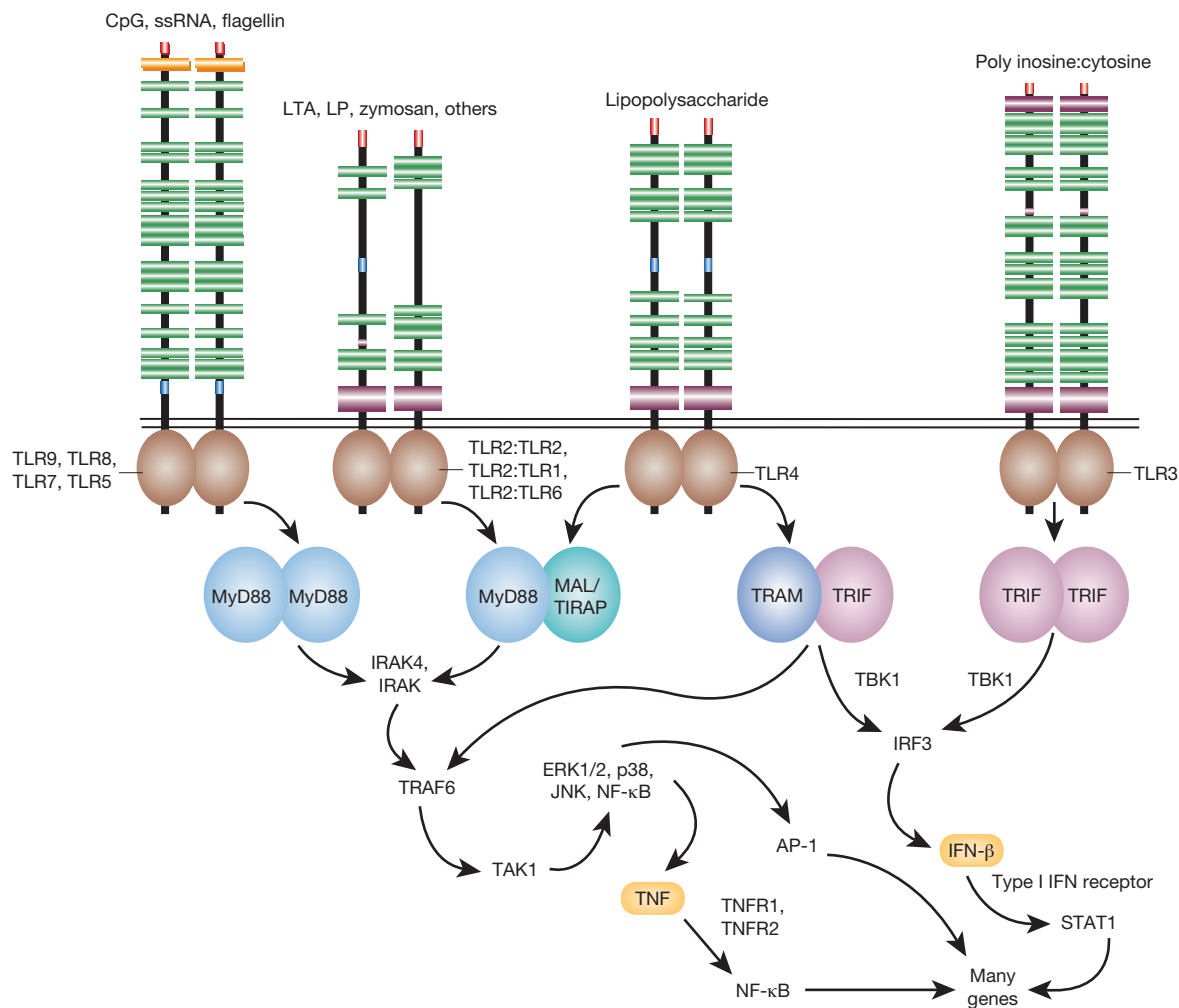


Figure 1 The principal relationships between the Toll-like receptors (TLRs), their adaptors, protein kinases that are linked to them, and downstream signalling effects. Most of the TLRs are believed to be homodimers, although heterodimers exist in the case of TLR1:TLR2 and TLR2:TLR6. The TIR (Toll/interleukin-1 receptor homologous region) adaptor proteins (MyD88, MAL/TIRAP, TRAM and TRIF) also appear to associate with one another and are drawn as homodimers or heterodimers. In humans, TLR10 (not shown) has unknown ligand specificity. Mice do not express TLR10, but do express TLR11, TLR12 and TLR13 (which are absent in humans; not shown). The ligands for these receptors remain unknown. Mouse TLR7 and human TLR8 are specific for single-stranded (ss)RNA. The corresponding orthologues have unknown specificity. IFN- β , interferon- β ; LP, lipopeptide; LTA, lipoteichoic acid; MAL/TIRAP, MyD88-adaptor-like/TIR-associated protein; MyD88, myeloid differentiation factor 88; TNF, tumour-necrosis factor; TRIF, Toll-receptor-associated activator of interferon; TRAM, Toll-receptor-associated molecule.

both to model inflammation and to assist in immunization. The same stimuli have also been used to model sepsis — the syndrome of hypotension, coagulopathy and organ damage that occurs during overwhelming infection. Now that the molecular basis of the response to microbial stimuli is known, inflammation, modulation of the adaptive immune response and sepsis are among the most important practical challenges to which our current understanding of innate immune sensing might be applied.

Sterile inflammation and its relationship to TLR signalling

Inflammation causes a vast amount of human morbidity and mortality. It is observed not only in infection, but also in autoimmune diseases, in the aftermath of sterile trauma and in diseases that are not readily categorized at all — diseases that are not pathogenically dependent on autoantibodies or autoreactive T cells, but which certainly involve swelling, tissue injury and the influx of immunologically active cells. Diseases caused by sterile inflammation (for example, rheumatoid arthritis, ankylosing spondylitis and psoriasis) pose some of the greatest practical and intellectual challenges in immunology and in medicine, leaving few disciplines untouched.

Recently, diseases that were not classically regarded as inflammatory (for example, atherosclerosis and Alzheimer's disease) have been reassessed, and are widely considered as such today.

It has been supposed that many types of sterile inflammation use the very same proteins and pathways as those traversed in the course of an infection. Is this the case? The question is an immensely important one, because it could direct us to new approaches to the treatment of these diseases, based on comprehensive 'upstream' blockade of signalling rather than interdiction of individual cytokines. The simple answer is that we do not know for sure. But arguing in the affirmative, many of the same cytokine mediators are produced in both sterile and infectious inflammation, and in some specific instances, these mediators have pathogenic importance in each process. By far the best example of a shared end point is tumour-necrosis factor (TNF).

Macrophage TNF production was the principal effect of LPS signalling that was examined in the positional cloning of the *Lps* locus. It was chosen because LPS was known to induce TNF²³, and in turn, TNF was known to mediate LPS-induced shock²⁴. Other cytokines, including interferon- γ ²⁵, interferon- β ²⁶ and IL-12 (ref. 27), have been found to contribute to LPS-induced shock as well, and two of

these cytokines (interferon- β and IL-12) are 'primary' mediators in the sense that they are produced by monocyte macrophages in direct response to LPS.

Several primary cytokine mediators of LPS-induced shock might serve as mediators of sterile inflammatory diseases. The case is entirely clear for one of them. TNF blockade is very effective in the treatment of rheumatoid arthritis²⁸, ankylosing spondylitis²⁹, Crohn's disease³⁰ and psoriasis³¹. The reason that TNF is so important in the pathogenesis of these diseases is unclear, and only *post hoc* hypotheses have been offered to account for the efficacy of blockade. TNF signals through an evolutionarily conserved pathway and is a point of connection between two arms of the mammalian innate immune response. It is a second-wave activator of nuclear factor (NF)- κ B, capable of spreading the alarm sounded by TLRs to virtually all cell types in the body. In *Drosophila*, the Imd (immunodeficiency) pathway³² incorporates signalling elements that are used by the p55 and p75 TNF receptors (TNFRs), but has no known connection to the Toll pathway. In mammals, on the other hand, the TLR and TNF pathways are fused to one another by TNF itself (Fig. 2). TNF blockade severs this connection, isolating a major effector system from the primary response apparatus.

Given these observations, we could attempt to work backwards to establish the cause of TNF production in these diseases. The same sort of inquiry was taken with respect to attacin in *Hyalophora*

*cecropia*³³, as well as dipterin and drosomycin in *Drosophila*^{2,34,35}, and it emerged that there was but one pathway for drosomycin induction, proceeding by way of Toll^{2,36}. Regrettably, the situation in mammals is more complex, and potentially, there are quite a number of ways to generate TNF. To begin with, we do not know the cellular source of TNF in such diseases. And whatever the source, at least some TNF might arise from several pathways.

In myeloid cells, NF- κ B activation is a minimal condition for TNF biosynthesis³⁷, but not the sole condition, because de-repression of the translational blockade imposed by the UA-rich sequence in the TNF 3' untranslated region is also necessary, at least for efficient protein synthesis³⁸. Several members of the TNFR superfamily are capable of activating NF- κ B. This includes both of the TNFRs themselves, operating through quite separate signalling intermediates^{39,40}. But if we are looking for a cause of aberrant TNF synthesis *ab initio*, these pathways can be discounted. Other receptors of the TNFR family may also contribute to NF- κ B activation, and although none of these is known to be a strong TNF inducer, nor can they be dismissed as sources *a priori*. Transforming growth factor- β (TGF- β) can activate NF- κ B^{41–44}, perhaps by stimulating TGF- β -activated kinase 1 (TAK1). And IL-1, which signals through TIR domain receptors, is capable of activating NF- κ B in macrophages as well.

In lymphoid cells, the story is quite different. The T-cell receptor can activate the calcineurin pathway, yielding TNF synthesis, and

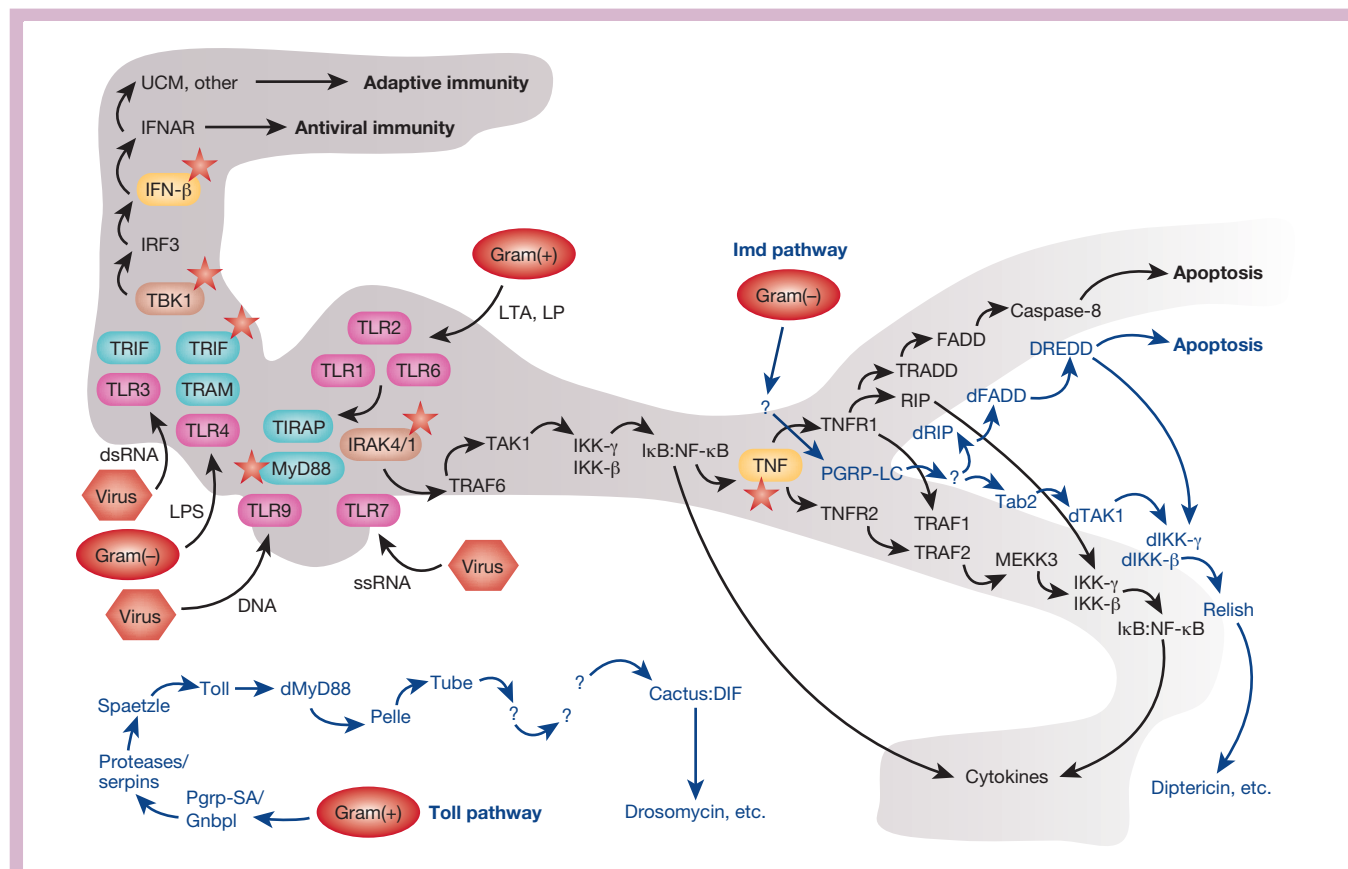


Figure 2 A global view of the signalling pathways activated by mammalian Toll-like receptors (TLRs) (grey shaded area), with reference to homologous pathways in *Drosophila* (blue lettering), reveals strategic targets for interventional blockade (red stars). Some of these targets have already been used. TLRs, shown in pink, can be blocked selectively and individually by antagonists such as E5531. Such an approach would be expected to be very specific, and would not cause strong immunosuppression. For some purposes, much broader inhibition would be desirable. Adaptors (green) can be blocked pharmacologically, but kinases (brown) often prove to be more accessible targets. Key effector cytokines (yellow) are especially attractive targets because of the high specificity with which they can be blocked. Inhibition of tumour-necrosis factor (TNF) 'amputates' a major part of the ancestrally conserved innate immune response pathway, corresponding to the Imd (immunodeficiency) pathway in *Drosophila*, yet has acceptable immunosuppressive effects in mammals. The mammalian pathways are in continuity with one another in the sense that all pathways shown will be activated by any TLR stimulus. On the other hand, the *Drosophila* Imd and Toll pathways have become independent entities. dsRNA, double-stranded RNA; IFN- β , interferon- β ; IRAK, IL-1R-associated kinase; LP, lipopeptide; LPS, lipopolysaccharide; LTA, lipoteichoic acid; MyD88, myeloid differentiation factor 88; ssRNA, single-stranded RNA; TBK1, TANK-binding kinase 1; TIRAP, TIR-associated protein; TNF, tumour-necrosis factor; TRIF, Toll-receptor-associated activator of interferon; TRAF, TNF-receptor-associated factor; TRAM, Toll-receptor-associated molecule.

NF-AT, like NF- κ B, is believed to be important⁴⁵. Costimulatory pathways, including the CD28 pathway, can also bring about NF- κ B activation through activation of the MAGUK (membrane-associated guanylate kinase) protein Carma-1, and in turn protein kinase- θ ^{46–52}.

If we wish to assign culpability to one cell type or the other, the effect of drugs that inhibit calcineurin signalling may be considered. Cyclosporin is considered to be helpful in rheumatoid arthritis, but is not as effective as TNF neutralization, despite the fact that it eliminates the production of many lymphoid activation markers, including cytokines such as IL-2 and interferon- γ . This fact is consistent with the interpretation that much, or most, of the TNF that is produced in rheumatoid arthritis originates in myeloid cells, which are not susceptible to the inhibitory effects of cyclosporin on TNF synthesis, because a different pathway is used in these cells.

But we are left in a state of uncertainty. The TNF production that is witnessed during sterile inflammation *in vivo* might arise from TLR signalling, which is the most effective pathway for TNF production. The hypothesis is testable, and surely attractive. In the near future, drugs capable of blocking signalling through TLRs may well settle the question. But at present we simply do not know that it is so.

The concept of 'innate autoimmunity'

If we take a leap of faith and grant that TLR signalling contributes to sterile inflammation, the top question must then centre on the cause: what endogenous signals initiate TLR signalling? Any immune system must have a mechanism for tolerance to self, and the basis of innate immune tolerance is the failure of TLRs to recognize molecular components of the host, at least under normal circumstances. So it is possible to envision a failure of this type of tolerance, and to refer to 'innate autoimmunity'. Innate autoimmunity (Box 1) might take two forms, neither excluding the other. First, the production of TNF and other cytokines, whether generalized or confined to a specific tissue, would certainly have untoward effects, and would be pathogenically significant in its own right if not promptly controlled. Second, the adjuvant effect of microbes is ultimately mediated by TLRs, and in the event that TLRs signal aberrantly for a sustained period of time concomitant with antigen presentation, an adaptive immune response to host proteins might be expected.

A number of reports that TLRs — and particularly TLR4 — recognize endogenous molecules, including fibrinogen, heat-shock proteins and other polypeptides, have emerged over the past few years. But definitive measures for the exclusion of 'exogenous' inducers have not always been applied, and in no instance has such an endogenous protein been widely validated as a TLR ligand. On the other hand, the conclusion that some endogenous molecules signal through TLRs is almost inescapable, if only because sterile inflammation so closely resembles infectious inflammation. It is doubtful that all of the endogenous inducers activate TLRs directly (that is, without the intercession of adaptor molecules), but some of them may indeed.

One endogenous inducer of TLR responses is DNA, either of genomic or of mitochondrial origin. Noting that antibodies against deoxyribonucleoprotein complexes are abundant in systemic lupus erythematosus, Marshak-Rothstein and colleagues have posited that the adaptive immune response is sustained in autoimmunity by the adjuvant effect of endogenous nucleic acids^{53–55}. This leaves open the question of what is primary, but without doubt, host DNA exhibits innate immunostimulatory potential, CpG methylation notwithstanding.

DNA sensing is mediated by TLR9. What of the other TLRs? Although host DNA is fundamentally similar to microbial DNA, the host is not known to have molecules that resemble LPS or bacterial lipoproteins. But there is precedent for host proteins that clearly do engage the TLRs, and such proteins might act as conduits for TLR activation, binding other (presumably modified or dysfunctional) molecules of the host in turn. In 1990, CD14 was identified as a biologically relevant LPS receptor⁵⁶, and later it was shown to initiate the

Box 1

Innate versus adaptive autoimmunity

Diseases once simply designated as 'autoimmune' may now be seen as problems of adaptive immunity or innate immunity, insofar as each system has retained its own distinct sensing and effector mechanisms. Although interactions between the two systems blur the distinction, it is clear that innate immune responses can injure the host just as adaptive immune responses do. Sepsis, the purest example of host injury mediated by an innate immune response, is acute and is caused by exogenous stimuli; other innate immune diseases are chronic and seem to have purely endogenous causes.

Box 1 Table **Systems of adaptive and innate autoimmunity**

	Adaptive autoimmunity	Innate autoimmunity
Cells	T cells, B cells	Macrophages, natural killer cells
Effectors	Antibodies, T cells	Cytokines
Cause	Failure of peripheral or central T-cell tolerance	Aberrant sensor activation or failure of inhibitory mechanisms
Examples	Myasthenia gravis, systemic lupus erythematosus, type I diabetes, Hashimoto's thyroiditis	Rheumatoid arthritis, Crohn's disease, ankylosing spondylitis, psoriasis, sepsis

LPS signal through TLR4 (refs 3, 57). Indeed, physical interaction between CD14 and TLR4 has been demonstrated, and it is now supposed that a ternary complex incorporating CD14, MD-2 and TLR4 serves to activate LPS signalling. LPS-binding protein (LBP) is generally believed to associate more loosely with the complex, and to carry to the complex not only LPS, but also phosphatidylinositol and perhaps other endogenous lipids⁵⁸.

CD14 and LBP are not known to bind endogenous proteins. But a second intermediary between TLRs and the microbial world clearly does. Using a forward genetic screen, CD36 has been identified as a protein that is required for TLR2 recognition of di-acylated bacterial lipopeptides and lipoteichoic acid (K. Hoebe *et al.*, unpublished observation). Other data, both biochemical and genetic, have established that CD36 binds free fatty acids, thrombospondin^{59,60}, oxidized low-density lipoprotein⁶¹ and β -amyloid protein⁶². Some of these interactions may retard the activation of TLR2; others may encourage it. For example, mice that lack thrombospondin-1, a CD36-binding protein, develop severe and fatal pulmonary inflammation⁶³ similar to that observed in TGF- β -deficient mice, indicating that thrombospondin-1 has an anti-inflammatory function. This protein might serve to inhibit TLR2 activation by CD36.

The primary lesions in inflammatory diseases might thus affect proteins that modulate activation through TLRs. Even small molecules (for example, long-chain fatty acids) that can be generated by *de novo* synthesis or consumed in the diet are known to interact with CD36 and might modulate its activity in relation to TLR2 activation.

The innate–adaptive connection

Under many circumstances, the adaptive immune response is either too weak or too strong. New strategies for vaccination would also open many doors in medicine. It is widely believed that diverse (and presently unconquerable) diseases such as malaria and AIDS would yield to vaccination, if only a high titre of polyspecific antibody could somehow be maintained indefinitely. The same holds true for many other infectious diseases, and perhaps for certain types of neoplasia. Autoimmunity and allergy are the flip side of the same coin. Armed with a detailed understanding of the process by which adaptive immunity is activated, it is appealing to think that we could either enhance or suppress it for clinical gain.

Eighty years ago, Lewis and Loomis noted that in guinea pigs, mycobacterial infection during the administration of a separate antigen could enhance the antibody response to the latter — a phenomenon they termed 'allergic irritability'⁶⁴. Freund and McDermott later showed that dead microbes also exerted an adjuvant effect⁶⁵, and

isolated LPS was found to do so in 1955 (ref. 66). So it has long been clear that adaptive immune responses flourish in the presence of specific molecules of microbial origin.

Today we know a great deal more about how microbes are perceived and what molecular signals they evoke. It is conceivable that microbial adjuvants might one day be obsolete, replaced by TLR agonists or perhaps by TLR mimetics. Beyond this, many (and perhaps all) of the factors that mediate adjuvant responses (including the upregulation of costimulatory molecules that drive the adaptive immune response) are cytokines. Interferon- β has a particularly important part to play^{21,67–69}, although undoubtedly, specific NF- κ B-regulated cytokines contribute as well²¹. Conceivably, the TLR pathways could be bypassed entirely.

For the present, synthetic DNA oligonucleotides bearing unmethylated CpG motifs (TLR9 agonists) have been used for their adjuvant properties⁷⁰. So, too, have LPS partial structures (TLR4 agonists) with diminished toxicity. DNA oligonucleotides have also been coupled to protein antigens, with striking effects on the adaptive immune response⁷¹. On the other hand, it is reasonable to question whether the therapeutic index has been enhanced by the use of 'monoreceptor adjuvants', and recent work indicates that chronic TLR9 stimulation is not bereft of toxicity⁷².

A more modest goal is modulation of the adaptive immune response: for example, making it more 'T_H1-like' by means of TLR stimulation⁷³. CpG oligonucleotides have been proposed for the treatment of asthma on this basis⁷⁴, and for the treatment of other diseases as well. It must be admitted that 'T_H1' and 'T_H2' are still descriptive terms in large part. Differences in the signalling pathways that enforce T_H1–T_H2 differentiation, not to mention the ultimate cause of a T_H1 or T_H2 phenotype, remain largely obscure. But further work may lead to a detailed understanding of precisely how TIR domain signalling can influence these pathways.

TLR signalling may contribute to other types of adaptive immune response as well. Allograft rejection is one of these. Allografts from MyD88-deficient mice that differ at a minor histocompatibility locus have been shown to survive for longer than control allografts⁷⁵. Early studies suggest that allografts from double-mutant mice lacking both MyD88 and TRIF are still more viable, able to endure for prolonged periods of time even if a major histocompatibility complex (MHC) difference is present (D. McKay, personal communication). It remains to be seen whether the signals that activate TLR activity under these conditions are truly endogenous. Moreover, although profound immunodeficiency is evident in mice lacking both MyD88 and TRIF (K. Hoebe *et al.*, unpublished observation), some adaptive responses are preserved. The next challenge may therefore be to find the molecular pathways for TIR-independent adaptive immune activation.

If allograft rejection is indeed quelled by blockade of TIR domain signalling, it stands to reason that at least some forms of autoimmunity will be as well. A tissue subject to autoimmune attack is in most respects comparable to an allograft — recognized on the basis of specific antigens, attacked by adaptive immune response and perceived as foreign although it is not. The possibilities loom very large, but at present, they are only possibilities and only a few studies have been performed.

Sepsis and its treatment

Since the advent of antibiotics, little improvement has been seen in the prognosis of severe sepsis (a word that literally means 'rot' or 'putrefaction', but which in medical parlance denotes the derangement of host physiology observed as the result of a severe infection). Sepsis continues to kill hundreds of thousands of people each year. Although infection can generally be controlled with antibiotics, the systemic inflammatory response is quite another matter. Knowing which molecules activate the TLRs as we now do, and knowing in some detail how TLRs signal, we are in a far more powerful position to intervene in sepsis than ever before.

The intensity of the innate immune response has been calculated by natural selection, and reflects the fact that most encounters between host and pathogen are relatively small-scale. In the daily life of the host, a small number of organisms are more likely to be inoculated than a vast number, and a vigorous response within the microenvironment offers the best chance for eradication of the infection before it overwhelms the host. The downside of a vigorous response is observed on those rare occasions in which the inoculum is large, or when, because of chance or immunodeficiency, the infection is not quickly contained. On such occasions, a life-threatening systemic response to the pathogen may occur.

LPS was long taken as a molecule that could elicit most (and perhaps all) of the discernible effects of an authentic infection, including fever, shock and other features of sepsis. With the cloning of *Lps*, all cellular effects of LPS were traced to a single molecule³. Of course, other microbial molecules (for example, lipopeptides, lipoteichoic acid, unmethylated DNA and flagellin) also have LPS-like effects, but these effects have now been traced to other TLRs. As a point of logic, the TLRs initiate cellular responses to virtually all microbial signals, and in the absence of TLRs, most microbes elicit few if any responses at all. So these receptors must be assigned responsibility for sepsis, at least as it usually occurs in the mammalian host. This highly reductionist conclusion can be asserted despite the diversity of microbial inducers that are involved. It can be said, then, that we have a precise molecular understanding of the proximal cause of sepsis. And to some extent, we also know how to arrest the primary signals that are responsible for sepsis.

Expectations concerning the treatment of sepsis have remained relatively modest — perhaps rightly so, but perhaps not. In part, a sense of frustration has been engendered by numerous unsuccessful attempts at interventional therapy, ranging from steroids to anti-LPS antibodies to the use of IL-1R antagonists and anti-TNF. Indeed, only activated protein C⁷⁶ and low-dose steroids⁷⁷ are generally held to be effective. The lack of efficacy of specific anti-cytokine approaches in sepsis has led to pessimistic attitudes towards the problem as a whole.

The TLRs are a target of a new kind. They clearly mediate the recognition of microbes, and are therefore ultimately responsible for all aspects of the host response that we know as sepsis. Apart from the TLRs and their proximal signalling apparatus, there are no other targets of which blockade could be expected to yield comprehensive indifference to infection. It has been argued, with some justification, that 'the action is over' by the time an attempt to intervene in sepsis is made. Although it is not possible to refute this argument, and although the response of macrophages to TLR activators is very prompt *in vitro*, it is difficult to examine the question of timing for human infections acquired under natural conditions *in vivo*. The temptation to target the TLR signalling pathways in sepsis is therefore compelling, and no doubt it will ultimately be pursued.

An entirely different (although quite compatible) approach starts from the premise that infections would not usually cause sepsis if innate immune defence were not somehow flawed; hence, we should seek to find innate immunodeficiencies to identify individuals who are at risk. Some progress has already been made with this approach. For example, patients with systemic meningococcal disease, taken as a group, show a highly significant excess of low-frequency coding variation within the *TLR4* gene as compared with ethnically matched controls⁷⁸. Patients with a common nonsense error in *TLR5* appear to be more susceptible to infection by *Legionella pneumophila*⁷⁹. These studies offer some reason for hope that a 'manageable' number of genes may predict a substantial fraction of host susceptibility, at least for well-defined pathogens.

On the other hand, a broad search for genes required for an effective innate immune response has been undertaken using *N*-ethyl-*N*-nitrosourea (ENU) mutagenesis in mice⁵. Insofar as many such mutations have now been produced and mapped (although saturation of the genome is a distant prospect), it seems safe to estimate that hundreds of loci support innate immunity. A screening approach

aimed at the comprehensive prediction of susceptibility to sepsis will probably involve the implementation of technologies that have not yet been developed.

What should be targeted and when?

We cannot yet be dogmatic about what should be done; rather, we can only speak of what can be done. At present, it is possible to block signalling through individual TLRs (such as TLR4; ref. 80) and to disrupt TIR domain interactions⁸¹. And we can predict with confidence that it will also be possible to block kinases such as IRAK4 (IL-1R-associated kinase 4) and TBK1 (TANK-binding kinase 1). The potential benefits of blocking TLR signalling in sepsis and in chronic inflammation have already been discussed. But a price must be paid for paralyzing innate immunity in the attempt to quell inflammation, whether of infectious or non-infectious cause. Ideally (although perhaps not exclusively), one would like to pay this price under conditions in which inflammation: (1) is chronic (there is no point in blocking upstream events after they have already taken place if they will probably not be sustained or repeated); (2) is sterile (it is generally not wise to block innate immune defences when they are accomplishing something desirable); (3) uses the same pathways as those used by the innate immune system (otherwise, therapy is off-target); and (4) is either debilitating or life-threatening.

If these criteria are met, and if TLR pathways are indeed used in the disease of interest, the price of innate immune paralysis may be deemed acceptable, and certain discrete points of therapeutic intervention may be selected as relative 'bargains'. Some such bargains have already been identified and validated by germline mutations. In choosing targets, the following considerations apply.

The innate immune response has an 'hourglass' shape (Fig. 3). Bacteria are extremely complex at the molecular level, composed of thousands of proteins, large quantities of nucleic acids, and a great many types of small organic molecules. However, the host is oblivious to the vast majority of these molecules. Only a handful them (for example, LPS, lipopeptides, lipoteichoic acid, flagellin and unmethylated DNA) incite an innate immune response.

The microbial inducers just named provoke the bulk of the host innate immune response to a bacterium, and although assisted to their receptors by other proteins, including LBP, CD14 and CD36, their presence is sensed by as few as six of the TLRs. The signals that emanate from these six TLRs can be abolished, at least in large part and perhaps completely, by disrupting only two of five known TIR motif adaptor proteins (MyD88 and TRIF) that transduce signals from TLRs. So the whole of the innate immune signal traverses a narrow strait. Thereafter, it grows in complexity.

The signal is propagated by the serine kinase IRAK4 (ref. 82), and if LPS is the inciting stimulus, also by way of the TRAM-TRIF-TRAF6 (TNF-receptor-associated factor 6)^{14,83-85} and TRAM-TRIF-TBK1-IRF3 (interferon regulatory factor 3)^{14,83,84} pathways. TRAF6 coordinates the activation of still other protein kinases, which collectively lead to the activation of numerous transcription factors. Activated NF- κ B, AP-1 and IRF3 are responsible for most of the hundreds to thousands of transcriptional changes that ensue in the immediate aftermath of cell activation.

For viruses, sensing pathways appear to be very similar, although nucleic acids are of primary importance as inducers^{4,14}. The innate immune response is degenerate¹⁴, and to a very large extent, one size fits all.

Some of the transcripts and the proteins they encode are clearly more important to the inflammatory response than others, and it should not be forgotten that post-transcriptional regulatory events are also set in motion and contribute to the altered state of the cell in important ways. Among the primary cytokines produced in response to TLR activation, TNF, IL-12 and the type I interferons are of key importance for the induction of further innate immune processes and also for activation of adaptive immunity. One might think that the redundancy of cytokine effects would make intervention at the

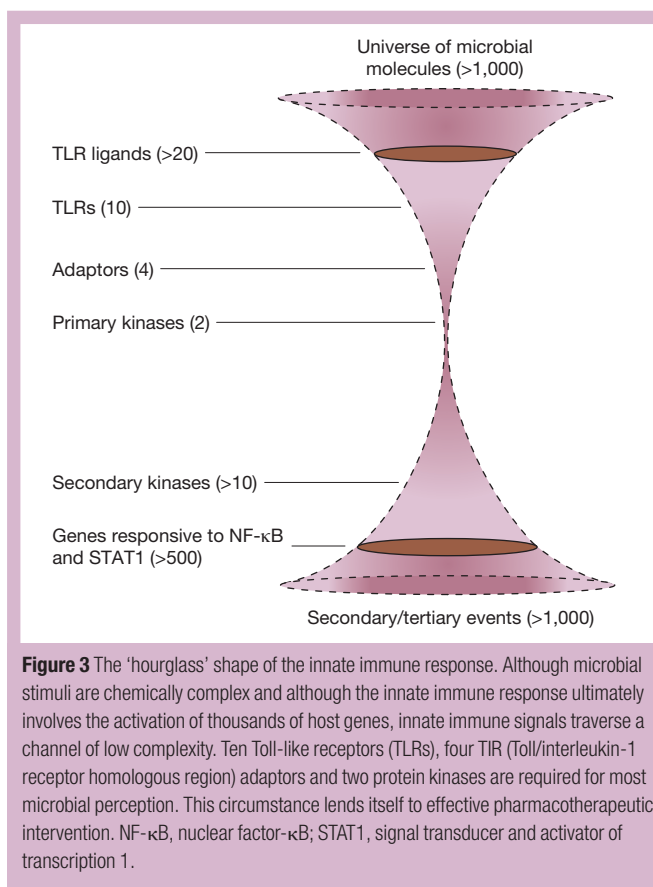


Figure 3 The 'hourglass' shape of the innate immune response. Although microbial stimuli are chemically complex and although the innate immune response ultimately involves the activation of thousands of host genes, innate immune signals traverse a channel of low complexity. Ten Toll-like receptors (TLRs), four TIR (Toll/interleukin-1 receptor homologous region) adaptors and two protein kinases are required for most microbial perception. This circumstance lends itself to effective pharmacotherapeutic intervention. NF- κ B, nuclear factor- κ B; STAT1, signal transducer and activator of transcription 1.

cytokine level a poor prospect. But surprisingly, neutralization of individual cytokines — particularly TNF — has been highly effective in the treatment of selected non-infectious inflammatory diseases. How much more effective might we be if intervention were taken to a higher level of the signalling pathway?

It has been almost exactly six years since the mammalian TLRs were recognized as proteins that allow awareness of infection. With this, the molecular basis of our inherited ability to recognize conserved molecules of microbial origin was revealed, opening a window into the activation of the innate immune system and all that follows. Time will attest to the wisdom of blocking or stimulating TLR signals in one clinical setting or another. For the moment, we only know that opportunities have been created where none existed a short time ago. □

doi:10.1038/nature02761

- Anderson, K. V., Bokla, L. & Nusslein-Volhard, C. Establishment of dorsal-ventral polarity in the *Drosophila* embryo: the induction of polarity by the Toll gene product. *Cell* **42**, 791–798 (1985).
- Lemaitre, B., Nicolas, E., Michaut, L., Reichhart, J. M. & Hoffmann, J. A. The dorsoventral regulatory gene cassette *spätzle*/Toll/cactus controls the potent antifungal response in *Drosophila* adults. *Cell* **86**, 973–983 (1996).
- Poltorak, A. *et al.* Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in *Tlr4* gene. *Science* **282**, 2085–2088 (1998).
- Tabeta, K. *et al.* TLR9 and TLR3 as essential components of innate immune defense against mouse cytomegalovirus. *Proc. Natl Acad. Sci. USA* **101**, 3516–3521 (2004).
- Beutler, B., Hoebe, K., Georgel, P., Tabeta, K. & Du, X. Forward genetic dissection of afferent immunity: the role of TIR adapter proteins in innate and adaptive immune responses. *Adv. Exp. Med. Biol.* (in the press).
- Akira, S., Takeda, K. & Kaisho, T. Toll-like receptors: critical proteins linking innate and acquired immunity. *Nature Immunol.* **2**, 675–680 (2001).
- Campos, M. A. *et al.* Activation of Toll-like receptor-2 by glycosylphosphatidylinositol anchors from a protozoan parasite. *J. Immunol.* **167**, 416–423 (2001).
- Campos, M. A. *et al.* Impaired production of proinflammatory cytokines and host resistance to acute infection with *Trypanosoma cruzi* in mice lacking functional myeloid differentiation factor 88. *J. Immunol.* **172**, 1711–1718 (2004).
- Underhill, D. M. *et al.* The Toll-like receptor 2 is recruited to macrophage phagosomes and discriminates between pathogens. *Nature* **401**, 811–815 (1999).
- Takeuchi, O. *et al.* Differential roles of TLR2 and TLR4 in recognition of Gram-negative and Gram-positive bacterial cell wall components. *Immunity* **11**, 443–451 (1999).

11. Hayashi, F. *et al.* The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* **410**, 1099–1103 (2001).
12. Zhang, D. *et al.* A Toll-like receptor that prevents infection by uropathogenic bacteria. *Science* **303**, 1522–1526 (2004).
13. Meier, A. *et al.* Toll-like receptor (TLR) 2 and TLR4 are essential for *Aspergillus*-induced activation of murine macrophages. *Cell Microbiol.* **5**, 561–570 (2003).
14. Hoebe, K. *et al.* Identification of Lps2 as a key transducer of MyD88-independent TIR signalling. *Nature* **424**, 743–748 (2003).
15. Lund, J., Sato, A., Akira, S., Medzhitov, R. & Iwasaki, A. Toll-like receptor 9-mediated recognition of Herpes simplex virus-2 by plasmacytoid dendritic cells. *J. Exp. Med.* **198**, 513–520 (2003).
16. Heil, F. *et al.* Species-specific recognition of single-stranded RNA via Toll-like receptor 7 and 8. *Science* **303**, 1526–1529 (2004).
17. Diebold, S. S., Kaisho, T., Hemmi, H., Akira, S. & Reis e Sousa, C. Innate antiviral responses by means of TLR7-mediated recognition of single-stranded RNA. *Science* **303**, 1529–1531 (2004).
18. Lund, J. M. *et al.* Recognition of single-stranded RNA viruses by Toll-like receptor 7. *Proc. Natl Acad. Sci. USA* **101**, 5598–5603 (2004).
19. Chamallard, M., Girardin, S. E., Viala, J. & Philpott, D. J. Nods, Nalps and Naip: intracellular regulators of bacterial-induced inflammation. *Cell Microbiol.* **5**, 581–592 (2003).
20. Brown, M. G. *et al.* Vital involvement of a natural killer cell activation receptor in resistance to viral infection. *Science* **292**, 934–937 (2001).
21. Hoebe, K. *et al.* Upregulation of costimulatory molecules induced by lipopolysaccharide and double-stranded RNA occurs by Trif-dependent and Trif-independent pathways. *Nature Immunol.* **4**, 1223–1229 (2003).
22. Schnare, M. *et al.* Toll-like receptors control activation of adaptive immune responses. *Nature Immunol.* **2**, 947–950 (2001).
23. Beutler, B. *et al.* Identity of tumour necrosis factor and the macrophage-secreted factor cachectin. *Nature* **316**, 552–554 (1985).
24. Beutler, B., Milsark, I. W. & Cerami, A. Passive immunization against cachectin/tumour necrosis factor (TNF) protects mice from the lethal effect of endotoxin. *Science* **229**, 869–871 (1985).
25. Car, B. D. *et al.* Interferon γ receptor deficient mice are resistant to endotoxic shock. *J. Exp. Med.* **179**, 1437–1444 (1994).
26. Karagiosoff, M. *et al.* Central role for type I interferons and Tyk2 in lipopolysaccharide-induced endotoxin shock. *Nature Immunol.* **4**, 471–477 (2003).
27. Wysocka, M. *et al.* Interleukin-12 is required for interferon- γ production and lethality in lipopolysaccharide-induced shock in mice. *Eur. J. Immunol.* **25**, 672–676 (1995).
28. Elliott, M. J. *et al.* Randomised double-blind comparison of chimeric monoclonal antibody to tumour necrosis factor α (cA2) versus placebo in rheumatoid arthritis. *Lancet* **344**, 1105–1110 (1994).
29. Brandt, J. *et al.* Successful treatment of active ankylosing spondylitis with the anti-tumour necrosis factor α monoclonal antibody infliximab. *Arthritis Rheum.* **43**, 1346–1352 (2000).
30. Van Dullemen, H. M. *et al.* Treatment of Crohn's disease with anti-tumour necrosis factor chimeric monoclonal antibody (cA2). *Gastroenterology* **109**, 129–135 (1995).
31. Iyer, S., Yamauchi, P. & Lowe, N. J. Etanercept for severe psoriasis and psoriatic arthritis: observations on combination therapy. *Br. J. Dermatol.* **146**, 118–121 (2002).
32. Georgel, P. *et al.* *Drosophila* immune deficiency (IMD) is a death domain protein that activates antibacterial defense and can promote apoptosis. *Dev. Cell* **1**, 503–514 (2001).
33. Sun, S. C., Lindstrom, L., Lee, J. Y. & Faye, I. Structure and expression of the attacin genes in *Hyalophora cecropia*. *Eur. J. Biochem.* **196**, 247–254 (1991).
34. Reichhart, J. M. *et al.* Insect immunity: developmental and inducible activity of the *Drosophila* dipterin promoter. *EMBO J.* **11**, 1469–1477 (1992).
35. Georgel, P. *et al.* Insect immunity: the dipterin promoter contains multiple functional regulatory sequences homologous to mammalian acute-phase response elements. *Biochem. Biophys. Res. Commun.* **197**, 508–517 (1993).
36. Rutschmann, S. *et al.* The Rel protein DIF mediates the antifungal but not the antibacterial host defense in *Drosophila*. *Immunity* **12**, 569–580 (2000).
37. Shakhov, A. N., Collart, M. A., Vassalli, P., Nedospasov, S. A. & Jongeneel, C. V. κ B-type enhancers are involved in lipopolysaccharide-mediated transcriptional activation of the tumor necrosis factor α gene in primary macrophages. *J. Exp. Med.* **171**, 35–47 (1990).
38. Han, J., Brown, T. & Beutler, B. Endotoxin-responsive sequences control cachectin/TNF biosynthesis at the translational level. *J. Exp. Med.* **171**, 465–475 (1990).
39. Hsu, H. L., Shu, H. B., Pan, M. G. & Goeddel, D. V. TRADD–TRAF2 and TRADD–FADD interactions define two distinct TNF receptor 1 signal transduction pathways. *Cell* **84**, 299–308 (1996).
40. Rothe, M., Sarma, V., Dixit, V. W. & Goeddel, D. V. TRAF2-mediated activation of NF- κ B by TNF receptor 2 and CD40. *Science* **269**, 1424–1427 (1995).
41. Han, S. H., Yea, S. S., Jeon, Y. J., Yang, K. H. & Kaminski, N. E. Transforming growth factor- β 1 (TGF- β 1) promotes IL-2 mRNA expression through the up-regulation of NF- κ B, AP-1 and NF-AT in EL4 cells. *J. Pharmacol. Exp. Ther.* **287**, 1105–1112 (1998).
42. Arsur, M. *et al.* Transient activation of NF- κ B through a TAK1/IKK kinase pathway by TGF- β 1 inhibits AP-1/SMAD signaling and apoptosis: implications in liver tumor formation. *Oncogene* **22**, 412–425 (2003).
43. Sakurai, H., Chiba, H., Miyoshi, H., Sugita, T. & Toriumi, W. I κ B kinases phosphorylate NF- κ B p65 subunit on serine 536 in the transactivation domain. *J. Biol. Chem.* **274**, 30353–30356 (1999).
44. Sakurai, H., Miyoshi, H., Toriumi, W. & Sugita, T. Functional interactions of transforming growth factor β -activated kinase 1 with I κ B kinases to stimulate NF- κ B activation. *J. Biol. Chem.* **274**, 10641–10648 (1999).
45. Goldfeld, A. E. *et al.* Calcineurin mediates human tumor necrosis factor α gene induction in stimulated T and B cells. *J. Exp. Med.* **180**, 763–768 (1994).
46. Verweij, C. L., Geerts, M. & Aarden, L. A. Activation of interleukin-2 gene transcription via the T-cell surface molecule CD28 is mediated through an NF- κ B-like response element. *J. Biol. Chem.* **266**, 14179–14182 (1991).
47. Harhaj, E. W., Maggirwar, S. B., Good, L. & Sun, S. C. CD28 mediates a potent costimulatory signal for rapid degradation of I κ B β which is associated with accelerated activation of various NF- κ B/Rel heterodimers. *Mol. Cell. Biol.* **16**, 6736–6743 (1996).
48. Hehner, S. P. *et al.* Mixed-lineage kinase 3 delivers CD3/CD28-derived signals into the I κ B kinase complex. *Mol. Cell. Biol.* **20**, 2556–2568 (2000).
49. Lin, X., O'Mahony, A., Mu, Y., Geleziunas, R. & Greene, W. C. Protein kinase C- θ participates in NF- κ B activation induced by CD3–CD28 costimulation through selective activation of I κ B kinase β . *Mol. Cell. Biol.* **20**, 2933–2940 (2000).
50. Jun, J. E. *et al.* Identifying the MAGUK protein Carma-1 as a central regulator of humoral immune responses and atopy by genome-wide mouse mutagenesis. *Immunity* **18**, 751–762 (2003).
51. Marinari, B., Costanzo, A., Marzano, V., Piccolella, E. & Tosto, L. CD28 delivers a unique signal leading to the selective recruitment of RelA and p52 NF- κ B subunits on IL-8 and Bcl-xL gene promoters. *Proc. Natl Acad. Sci. USA* **101**, 6098–6103 (2004).
52. Wang, D. *et al.* CD3/CD28 costimulation-induced NF- κ B activation is mediated by recruitment of protein kinase C- θ , Bcl10, and I κ B kinase β to the immunological synapse through CARMA1. *Mol. Cell. Biol.* **24**, 164–171 (2004).
53. Leadbetter, E. A. *et al.* Chromatin-IgG complexes activate B cells by dual engagement of IgM and Toll-like receptors. *Nature* **416**, 603–607 (2002).
54. Viglianti, G. A. *et al.* Activation of autoreactive B cells by CpG dsDNA. *Immunity* **19**, 837–847 (2003).
55. Leadbetter, E. A., Rifkin, I. R. & Marshak-Rothstein, A. Toll-like receptors and activation of autoreactive B cells. *Curr. Dir. Autoimmun.* **6**, 105–122 (2003).
56. Wright, S. D., Ramos, R. A., Tobias, P. S., Ulevitch, R. J. & Mathison, J. C. CD14, a receptor for complexes of lipopolysaccharide (LPS) and LPS binding protein. *Science* **249**, 1431–1433 (1990).
57. Poltorak, A., Ricciardi-Castagnoli, P., Citterio, A. & Beutler, B. Physical contact between LPS and Tlr4 revealed by genetic complementation. *Proc. Natl Acad. Sci. USA* **97**, 2163–2167 (2000).
58. Wang, P. Y. & Munford, R. S. CD14-dependent internalization and metabolism of extracellular phosphatidylinositol by monocytes. *J. Biol. Chem.* **274**, 23235–23241 (1999).
59. Asch, A. S., Barnwell, J., Silverstein, R. L. & Nachman, R. L. Isolation of the thrombospondin membrane receptor 3. *J. Clin. Invest.* **79**, 1054–1061 (1987).
60. Dawson, D. W. *et al.* CD36 mediates the *in vitro* inhibitory effects of thrombospondin-1 on endothelial cells. *J. Cell Biol.* **138**, 707–717 (1997).
61. Endemann, G. *et al.* CD36 is a receptor for oxidized low density lipoprotein-1. *J. Biol. Chem.* **268**, 11811–11816 (1993).
62. El Khoury, J. B. *et al.* CD36 mediates the innate host response to β -amyloid 1. *J. Exp. Med.* **197**, 1657–1666 (2003).
63. Crawford, S. E. *et al.* Thrombospondin-1 is a major activator of TGF- β 1 *in vivo*. *Cell* **93**, 1159–1170 (1998).
64. Lewis, P. A. & Loomis, D. The formation of anti-sheep hemolytic amboceptor in the normal and tuberculous guinea pig. *J. Exp. Med.* **40**, 503 (1924).
65. Freund, J. & McDermott, K. Sensitization to horse serum by means of adjuvants. *Proc. Soc. Exp. Biol. Med.* **49**, 548–553 (1942).
66. Condie, R. M., Zak, S. J. & Good, R. A. Effect of meningococcal endotoxin on the immune response. *Proc. Soc. Exp. Biol. Med.* **90**, 355–360 (1955).
67. Le Bon, A. *et al.* Type I interferons potentially enhance humoral immunity and can promote isotype switching by stimulating dendritic cells *in vivo*. *Immunity* **14**, 461–470 (2001).
68. Le Bon, A. & Tough, D. F. Links between innate and adaptive immunity via type I interferon. *Curr. Opin. Immunol.* **14**, 432–436 (2002).
69. Honda, K. *et al.* Selective contribution of IFN- α / β signaling to the maturation of dendritic cells induced by double-stranded RNA or viral infection. *Proc. Natl Acad. Sci. USA* **100**, 10872–10877 (2003).
70. Krieg, A. M. Antitumor applications of stimulating toll-like receptor 9 with CpG oligodeoxynucleotides. *Curr. Oncol. Rep.* **6**, 88–95 (2004).
71. Maurer, T. *et al.* CpG-DNA aided cross-presentation of soluble antigens by dendritic cells. *Eur. J. Immunol.* **32**, 2356–2364 (2002).
72. Heikenwalder, M. *et al.* Lymphoid follicle destruction and immunosuppression after repeated CpG oligodeoxynucleotide administration. *Nature Med.* **10**, 187–192 (2004).
73. O'Neill, L. A. Therapeutic targeting of Toll-like receptors for inflammatory and infectious diseases. *Curr. Opin. Pharmacol.* **3**, 396–403 (2003).
74. Broide, D. & Raz, E. DNA-Based immunization for asthma. *Int. Arch. Allergy Immunol.* **118**, 453–456 (1999).
75. Goldstein, D. R., Tesar, B. M., Akira, S. & Lakkis, F. G. Critical role of the Toll-like receptor signal adaptor protein MyD88 in acute allograft rejection. *J. Clin. Invest.* **111**, 1571–1578 (2003).
76. Bernard, G. R. *et al.* Efficacy and safety of recombinant human activated protein C for severe sepsis. *N. Engl. J. Med.* **344**, 699–709 (2001).
77. Annane, D. *et al.* Effect of treatment with low doses of hydrocortisone and fludrocortisone on mortality in patients with septic shock. *J. Am. Med. Assoc.* **288**, 862–871 (2002).
78. Smirnova, I. *et al.* Assay of locus-specific genetic load implicates rare Toll-like receptor 4 mutations in meningococcal susceptibility. *Proc. Natl Acad. Sci. USA* **100**, 6075–6080 (2003).
79. Hawn, T. R. *et al.* A common dominant TLR5 stop codon polymorphism abolishes flagellin signaling and is associated with susceptibility to legionnaires' disease. *J. Exp. Med.* **198**, 1563–1572 (2003).
80. Kawata, T. *et al.* E5531, a synthetic non-toxic lipid A derivative blocks the immunobiological activities of lipopolysaccharide. *Br. J. Pharmacol.* **127**, 853–862 (1999).
81. Bartfai, T. *et al.* A low molecular weight mimic of the Toll/IL-1 receptor/resistance domain inhibits IL-1 receptor-mediated responses. *Proc. Natl Acad. Sci. USA* **100**, 7971–7976 (2003).
82. Suzuki, N. *et al.* Severe impairment of interleukin-1 and Toll-like receptor signalling in mice lacking IRAK-4. *Nature* **416**, 750–756 (2002).
83. Yamamoto, M. *et al.* Role of adapter TRIF in the MyD88-independent Toll-like receptor signaling pathway. *Science* **301**, 640–643 (2003).
84. Yamamoto, M. *et al.* TRAM is specifically involved in the Toll-like receptor 4-mediated MyD88-independent signaling pathway. *Nature Immunol.* **4**, 1144–1150 (2003).
85. Jiang, Z. *et al.* Poly(I-C)-induced Toll-like receptor 3 (TLR3)-mediated activation of NF κ B and MAP kinase is through an interleukin-1 receptor-associated kinase (IRAK)-independent pathway employing the signaling components TLR3–TRAF6–TAK1–TAB2–PKR. *J. Biol. Chem.* **278**, 16713–16719 (2003).

Acknowledgements This work was supported by a grant from the National Institutes of Health.

Competing interests statement The author declares that he has no competing financial interests.

Translating cell biology *in vitro* to immunity *in vivo*

Marianne Boes* & Hidde L. Ploegh

Department of Pathology, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA

(e-mails: Marianne_boes@hms.harvard.edu; ploegh@hms.harvard.edu)

*Present address: Department of Dermatology, Brigham and Women's Hospital, Harvard Institutes of Medicine, Suite 671, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA

The elimination of pathogens and pathogen-infected cells initially rests on the rapid deployment of innate immune defences. Should these defences fail, it is the lymphocytes — T cells and B cells — with their antigen-specific receptors that must rise to the task of providing adaptive immunity. Technological advances are now allowing immunologists to correlate data obtained *in vitro* with *in vivo* functions. A better understanding of T-cell activation *in vivo* could lead to more effective strategies for the treatment and prevention of infectious and autoimmune diseases.

Pathogens are always on the lookout for a host to sustain their life cycle and are an important cause of disease in humans. Because pathogens have a strong preference for particular tissue sites, interacting with or infecting only certain cell types, there is still a paucity of appropriate animal models for the study of human infectious disease. Even where animal models are available, key aspects of host–pathogen interaction cannot usually be observed directly and in real time. With the obvious exception of death or survival (the natural end points), host–pathogen interactions are assessed by measurement of surrogate parameters, such as the number of antigen-specific T cells generated, or the types and levels of cytokines produced. Now that the complete genome sequences of the main human pathogens are available, there is potential to use this information to explore microbial countermeasures that frustrate immune recognition at all levels. New therapeutic principles may then follow.

Since the discovery of the T-cell receptor (TCR) for antigen and the TCR's associated proteins, the exact manner in which this complex array of surface proteins transduces signals has been studied mainly in isolated T cells. The ability to observe the behaviour of T cells and antigen-presenting cells (APCs) in their natural habitat — for example, in lymph nodes or other tissues — is a recent addition to the immunologist's toolkit and has not been fully explored. The importance of an appropriate microenvironment^{1,2} and of signals derived from components of the innate immune system to achieve T-cell activation³ is clear. Here we discuss some of the successive stages of T-cell activation from a cell-biological perspective, highlighting opportunities for translating some of the cell-biological concepts to applications *in vivo*. The design of effective vaccination strategies and the prospects of immune intervention more generally should benefit from a better understanding of how the considerable body of *in vitro* data can apply to an *in vivo* setting.

Subversion of antigen presentation

The evolutionary success of viruses depends on their ability to disarm host defences, or to produce progeny in such overwhelming numbers that innate defences are inadequate and transfer of the virus to another host is a likely outcome. The possible countermeasures are many^{4,5}. Error-prone reverse transcription may allow human immunodeficiency virus

(HIV) to accumulate mutations faster than the adaptive immune system can effectively deal with the antigenically new variants that arise. Viruses with segmented genomes enjoy flexibility in their genetic make-up through a reassortment of the individual segments when the opportunity arises — for example, when a host is infected with two different yet related viruses, such as swine and avian strains of flu. The larger viruses, such as poxviruses and herpesviruses, probably do not have the luxury of simply mutating the gene that encodes a viral antigen. Instead they rely on interference with the cytokine and chemokine networks; they may disarm components of the complement system or interfere with antigen presentation⁶. Many of these arguments apply with equal force to bacterial pathogens⁷. Moreover, Gram-negative bacteria temporarily escape the host's immune system by using a so-called type III secretion system, which mediates translocation of pathogen-derived proteins across the eukaryotic cell membrane. *Yersinia enterocolitica* and *Salmonella typhimurium* deliver a diverse set of virulence factors directly into the cytosol to resist phagocytosis, a requirement for sustained extracellular bacterial replication, as discussed more fully in the review in this issue by Merrell and Falkow (page 250).

The presentation of peptides by major histocompatibility complex (MHC) molecules is essential for an effective and potent immune response, so it is not surprising that examples of interference of class I MHC-restricted antigen presentation are manifold. These include the adjustment of viral antigen synthesis, inhibition of peptide-antigen delivery to the class I MHC products, intracellular detention of newly assembled class I MHC proteins, or even physical destruction of MHC products by re-routing them either to lysosomes or to the cytosol for proteasomal degradation⁸. Too little is known at present about the lifestyle of many of these pathogens to state with precision which strategy is used *in vivo*. A large DNA virus such as human cytomegalovirus may encode over 200 gene products, some 60 of which may be dispensable for virus growth in tissue culture⁹ and are therefore probably determinants of the lifestyle of the virus. It is only recently that bacterial artificial chromosome (BAC)-based technology has become available¹⁰ to rapidly generate collections of deletion and insertion mutants in these large viral genomes. For some of these genes, their function can be determined only in the context of an *in vivo* infectious model, and then only if the

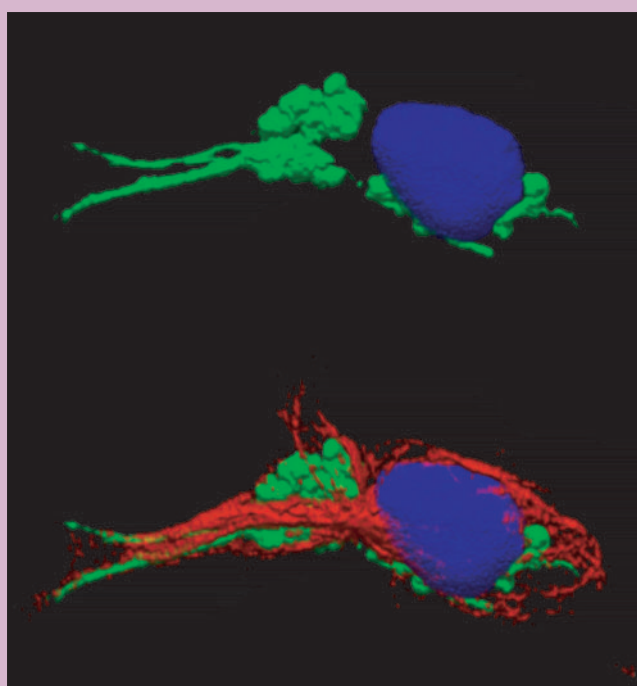


Figure 1 Dendritic cells rearrange their class II major histocompatibility complex (MHC)-positive endosomal compartments in a microtubule-dependent manner upon T-cell ligation. Bone-marrow-derived dendritic cells were grown from mice that express a class II MHC-eGFP (enhanced green fluorescent protein) fusion protein in place of endogenous class II MHC. Class II MHC-containing compartments form tubular structures that are polarized towards antigen-specific T cells. Tubulation was induced by the addition of T hybridoma cells specific for hen egg lysozyme (HEL) peptide^{46–61} to HEL-loaded dendritic cells. In the upper and lower panels, the nucleus is shown in blue (stained with Hoechst 33258); class II MHC-eGFP is in green. Dendritic cells were fixed, permeabilized and stained with antibody to α -tubulin, a key constituent of microtubules. We used a rat anti-mouse Cy3-labelled secondary antibody for detection (red in the lower panel). Reproduced from ref. 27.

appropriate parameters of innate or adaptive/acquired immunity are examined. The resurgent interest in highly pathogenic viruses that could be used for nefarious purposes, such as smallpox, underscores the need for a more thorough analysis *in vivo* of T-cell activation, including the steps immediately preceding it: acquisition of antigen by APCs, the migration of the antigen-laden APCs to the draining lymph node, and their encounter with a naive T cell. *In vivo*, each of these steps presents potential targets for intervention by pathogen-encoded genes.

Transport of antigen

APCs in peripheral immune organs are quiescent sentinels, activated when a pathogen is detected by Toll-like receptors (TLRs) (see review in this issue by Beutler, page 257), scavenger receptors or other surface structures. In particular, dendritic cells continuously survey their surroundings by fluid-phase uptake and endocytosis. Internalization of microbe-derived material induces their differentiation into potent APCs.

Only in lymphoid tissues does priming of naive T cells occur, because naive T cells are usually excluded from non-lymphoid tissues. For example, in the skin, tissue-derived antigens arrive in draining lymph nodes in two ways. During the first 24 hours, transport occurs mainly in free (not in cell-associated) form, in interstitial fluid delivered to the lymph node through afferent lymph vessels. Components of the complement system, in synergy with natural antibodies, help to prepare such antigens for efficient acquisition by the dendritic cell. In the lymph nodes, immature dendritic cells, which form class II MHC-peptide complexes that are used to prime naive T cells,

internalize the first wave of antigen. Thereafter, antigen-laden dermal dendritic cells arrive in the draining lymph node, having acquired their antigenic cargo in the periphery. These skin-derived dendritic cells, having already formed class II MHC-peptide complexes *en route* to their destination, display their MHC-peptide complexes to T cells in the lymph node¹¹. The relevance of these seemingly parallel pathways of antigen presentation must be sought in the heterogeneity of dendritic cells^{2,12}, which may vary in maturation stage, in their proteolytic machinery, and in their expression of costimulatory molecules. Costimulatory molecules may provide stimulatory or inhibitory signals that together control T-cell activation. The difference in the proteolytic capacity of the immigrant versus node-resident dendritic cells would generate class II MHC-peptide complexes that are not necessarily identical, although they derive from the same protein antigen. Different T-cell clones would therefore be activated, leading to more extensive T-cell activation, perhaps including an interconnected network of regulatory T cells.

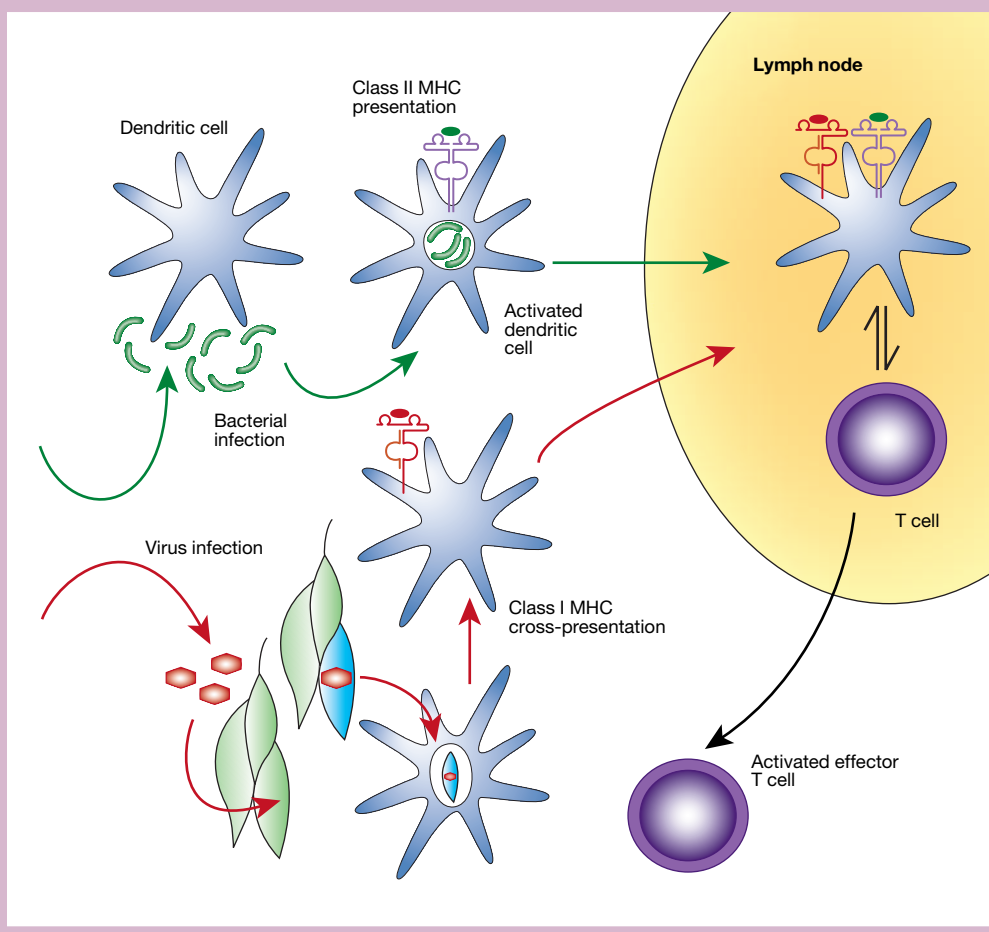
Trafficking of MHC molecules

Proper display of MHC products is a prerequisite for T-cell activation. The efficacy with which a protein antigen is converted to an MHC-peptide complex is not known with certainty, but estimates for a relatively well-studied antigen such as ovalbumin suggest that several thousand protein molecules must be destroyed to generate a single MHC-peptide complex¹³. This has made it technically difficult to track in live cells the conversion of an individual antigen molecule into its corresponding MHC-peptide complex. Perhaps antigen presentation on CD1 molecules, designed to present lipids to T cells¹⁴, could be exploited to examine the fate of synthetic fluorescently labelled glycolipids added exogenously or incorporated into the envelope of bacteria offered to an APC. Although a challenge from the synthetic perspective, the fact that ligands for CD1 are not template-specified, unlike ligands for the classical MHC products, affords possibilities for a chemical-biological approach to the question of how to visualize the ligand for a TCR in a living cell, or even *in vivo*.

Interactions between T cells and APCs can be short-lived, but in many cases contacts between antigen-specific T cells and APCs persist for hours¹⁵. If antigen were in limited supply, such prolonged interaction would imply serial engagement of TCRs by a limited number of MHC-peptide complexes¹⁶. The ability of an APC to carefully regulate this display is probably essential for allowing activation of naive T cells.

Immature dendritic cells are characterized by a paucity of costimulatory molecules, an abundance of class II products in intracellular compartments, and relatively modest quantities of class II MHC molecules displayed at the cell surface. Acquisition of antigen by endocytosis or phagocytosis, under most circumstances in which pathogens are present, will allow the dendritic cells to perceive signals through TLRs. The initial engagement of TLRs initiates a programme of differentiation or maturation^{17–19} that includes adjustment of endosomal pH and of active protease content^{20,21}. This increases the ability of the dendritic cell to process the newly acquired antigen and to load the resulting peptides onto class II MHC products. Specialized phagocytes such as macrophages internalize bacteria or apoptotic cells as a normal part of host defence²². Uptake of particulates induces phagosome formation, which by fusion with endosomal/lysosomal vesicles generates antigenic cargo for MHC and CD1 molecules^{23,24}. The fusion of phagosomes with endosomal/lysosomal compartments depends on the presence of TLR ligands²⁵. Phagocytosis of apoptotic cells induces a constitutive mode of phagosome maturation — as characterized by features such as low pH and a strongly reducing and highly proteolytic environment — whereas the complete destruction of bacteria requires the activation of TLRs²⁵. The introduction of phagosome-derived antigenic cargo into loading compartments for class II MHC and CD1 therefore appears to be regulated in part by ligation of TLR ligands.

Figure 2 Presentation pathways of exogenously acquired antigens. Proteins acquired from the surroundings of the antigen-presenting cell can provide a source of peptide for presentation on class II major histocompatibility (MHC) molecules, and in the case of cross-presentation, on class I MHC molecules. Top, bacteria-derived antigens (green) are internalized by dendritic cells and displayed as class II MHC–peptide complexes to antigen-specific CD4-positive T cells in draining lymph nodes. Bottom, dendritic cells can acquire viral antigens (red) by internalization of virally infected tissue (fibroblasts, muscle). Virus-infected tissue cells are proteolysed by the dendritic cells and proteins are processed into class I MHC–peptide complexes to allow activation of naive CD8-positive T cells.



Polarization of class II trafficking in APCs

The generation of green fluorescent protein (GFP)-tagged MHC products using a retroviral delivery system or gene-replacement technology has made it possible to track MHC products in live, professional APCs^{26,27}. This strategy allows examination of the response of APCs to ligands for TLRs, such as lipopolysaccharide (LPS)²⁶, a constituent of the outer membrane of Gram-negative bacteria, and of the APC's response to antigen-specific T cells^{27–29}. Exposure of bone-marrow-derived dendritic cells to LPS results in profound alterations in the morphology of endocytic compartments that are positive for class II MHC. Tubular compartments are formed, possibly by recruitment of internal membrane vesicles from multivesicular bodies³⁰, which are rich in class II MHC molecules. Engagement of an antigen-loaded dendritic cell by a T cell of the appropriate specificity likewise induces the formation of tubular compartments polarized towards the interacting T cell²⁷, a reaction for which previous engagement of a TLR seems to be necessary²⁸ (Fig. 1). What is not clear is whether such responses occur *in vivo*. Both the retroviral and gene-replacement technologies hold considerable promise for the examination *in vivo* of the behaviour of other molecules that control T-cell activation and function. There are obvious limitations to these approaches, because proteins are tagged whose cytoplasmic domains do not easily tolerate such engineered fusions without compromising their ability to transduce signals or interact with partner proteins. Ideally, the functional complementation of a null allele created by gene targeting with the GFP-labelled version of the same gene product would provide the genetic argument necessary to validate this approach. For receptor behaviour in primary cells of the immune system, surprisingly few examples have been subjected to this test *in vivo*. The analysis of interactions between surface proteins and the underlying cytoskeletal elements in primary APCs and T cells would benefit from similarly engineered fusion proteins.

Microdomains

The debate about the nature or even the existence of lipid rafts in live cells is unlikely to subside any time soon^{31–34}. Discovered as membrane specializations that survive extraction in non-ionic detergents and are enriched in glycolipids and in proteins that carry fatty acyl modifications, the size and relevance of these rafts remain unclear. Munro³⁵ raised a number of objections and suggested further experiments to address both the occurrence and the relevance of such membrane specializations. Recent experiments involving fluorescence resonance energy transfer (FRET) measurements suggest a far more isotropic distribution of components that are believed to be raft-specific³⁶, and show that the complexes that comprise these rafts may be far smaller than was originally suggested. Although this is not the place to summarize all the arguments (for reviews, see refs 31, 35), the existence of membrane specializations is easy to accept. Cell biologists are familiar with the distinction between apical and basolateral domains of epithelial cells³⁷ and the pronounced polarization of cells of the nervous system³⁸. But lymphocytes are small and round; they move and change shape, and consequently it is difficult to pin down a cell at any given time without fixation.

Although many investigators use detergent solubility or flotation of extracted materials on sucrose gradients as criteria for localization of their favourite protein(s) to lipid rafts, there are of course other data supporting the existence of plasma membrane microdomains. Lymphocytes and dendritic cells are richly endowed with proteins that belong to the family of membrane proteins called tetraspanins (after their characteristic four transmembrane domains), such as CD63, CD81 and CD9 (ref. 39). Biochemical analyses show that these proteins tend to form complexes with themselves and with other members of the tetraspanin family⁴⁰, and that they recruit other membrane proteins to these tetraspanin webs to form microdomains. Subtle perturbation of membrane structure in dendritic cells abolishes their

ability to stimulate T cells, even when the display of the MHC–peptide ligands themselves is unaffected⁴¹. A combination of intravital imaging methods in conjunction with biophysical techniques such as FRET may help to define the existence and nature of membrane microdomains *in vivo*, both for T cells and for APCs.

Cross-presentation

Many viruses display pronounced tropism, infecting only certain cell types. Even if the virus fails to infect an APC directly, a successful immune response can nonetheless be mounted. This is provided that professional APCs can acquire dying cells whose death may be the consequence of the virus infection (for example, through direct cytopathic effects or by the induction of apoptosis). The ability of APCs to acquire such cell remnants by phagocytosis and then present processed materials derived from these cell remnants by means of class I MHC molecules is called ‘cross-presentation’⁴² (Fig. 2). Mice rendered transgenic for the poliovirus receptor sustain productive infection with poliovirus, a pathogen normally incapable of infecting mice. If these animals are lethally irradiated and reconstituted with bone marrow from donors that do not express the poliovirus receptor, a robust response involving CD8-positive T cells is nonetheless observed. This requires the presence of a functional TAP (transporter associated with antigen presentation) complex in the donor bone marrow cells used for reconstitution, and provides strong evidence that bone-marrow-derived APCs can indeed access the class I pathway of antigen presentation, even when not infected with the virus themselves⁴³. There are at least two possible means by which antigen could be transferred: proteolysed fragments of the antigen or MHC-bound peptides could be liberated from donor cells and then transferred. Escape of such small fragments from a phagosomal compartment might be relatively straightforward and could involve peptide-conducting channels such as TAP, Sec61 or others^{44–47}. Alternatively, whole proteins may be the predominant source of peptides. After transfer to the cytosol, such proteins would undergo ubiquitin-dependent proteasomal degradation. This mode of transfer presumably requires protein-conducting channels with an architecture somewhat different from that of channels used for cotranslational protein import into the endoplasmic reticulum (ER), unless complete unfolding of the protein substrate precedes translocation.

The ability of phagocytic cells to recruit ER membrane to the newly formed phagosome⁴⁸ suggests that the normal constituents of the ER, such as the TAP complex or the Sec61 translocon complex, may be involved in this mode of antigen presentation^{49,50} (Fig. 3). The possibilities for translocation of antigen from the phagosome are by no means limited to these known players, as new escape routes from the ER, and by inference, from phagosomes that have incorporated ER materials, continue to be discovered⁵¹.

If indeed this mode of acquisition is the most important means of delivering antigens for cross-presentation, it follows that those antigens that are abundantly present as proteins will be cross-presented more successfully than the already processed peptides present on donor MHC molecules themselves, or as has been argued, in a complex with members of the heat-shock protein family such as gp96 (ref. 52). New evidence now suggests that this is indeed the case. Elegant experiments show that signal-sequence-derived epitopes or other epitopes predicted to be short-lived are cross-presented poorly, but once grafted onto the body of a carrier protein as a more stable entity, their cross-presentation improves dramatically^{53–55}.

Cross-presentation must be important not only for the recognition of virus-infected cells, but also as an unavoidable consequence of lymphocyte development as well. In the thymus, where the vast majority of newly arising T cells never reach maturity and die by apoptosis, bone-marrow-derived cells in the thymus continuously remove cell remnants. Because such cross-presentation presumably occurs in the absence of an inflammatory stimulus, T cells exposed to these cross-presented materials will be instructed to die, or may never acquire full functional competence and instead become anergic. The

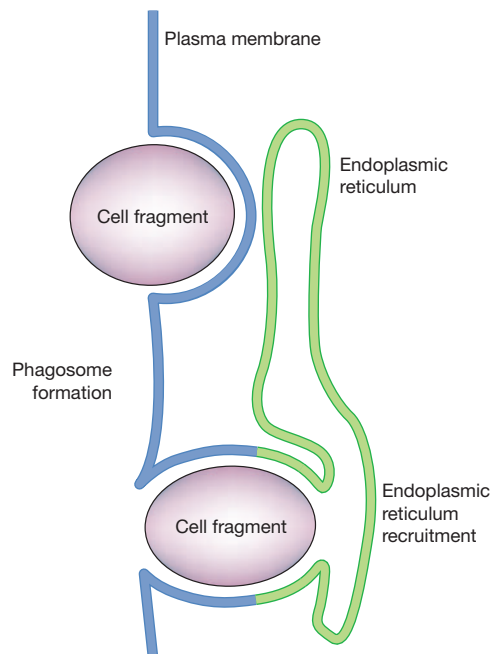


Figure 3 Fusion of phagosome with endoplasmic reticulum membrane. Phagocytic cells (macrophages and dendritic cells) can internalize cell fragments of significant size by recruiting membranes of the endoplasmic reticulum. Internalized exogenous antigens may therefore have immediate access to enzymes and transporters within the endoplasmic reticulum that are involved in class I major histocompatibility complex (MHC) presentation, facilitating cross-presentation.

details of cross-presentation are not yet understood, but experience with the more conventional trafficking pathways of MHC products will provide a useful backdrop.

Activation of T cells by APCs

Activation of T cells occurs when a professional APC displays the appropriate set of ligands to a naive T cell. These ligands must include a proper MHC–peptide complex, as well as the requisite sets of costimulatory molecules. Microbial products, such as cell wall constituents and nucleic acids, trigger members of the TLR family (see review in this issue by Beutler, page 257) and expression of costimulatory molecules. The importance of these surface-displayed costimulatory glycoproteins is apparent from inclusion of the appropriate blocking antibodies, which interfere with the interactions between APCs and T cells. Less clear is the relevance of alterations in cytoskeletal proteins, or in the detailed arrangement of components of the vacuolar system in both the APC and the T cell. Among the early changes seen in an activated T cell is the reorientation of the vacuolar apparatus^{56,57} so that release of cytokines, or the delivery of a lethal hit in the form of perforins and granzymes, will occur in a polarized manner, towards the interacting APC. Rearrangements of the cytoskeletal apparatus and of the microtubule network clearly occur, but how the directionality of these signals is controlled is not well understood. Ion gradients, maintained by controlling influx and efflux, are also under the influence of receptor-mediated events⁵⁸, but compared with the cells of the nervous system, much less is known about the role of ion channels in T-cell activation. Voltage-gated potassium channels do accumulate in the immunological synapse⁵⁹, a structure whose formation requires the ligation of TCRs and a variety of adhesion molecules (see below), but the functional consequences of such recruitment are unclear.

T-cell activation is assessed by monitoring antigen-specific engagement of the TCR, and by examining subsequent receptor-proximal

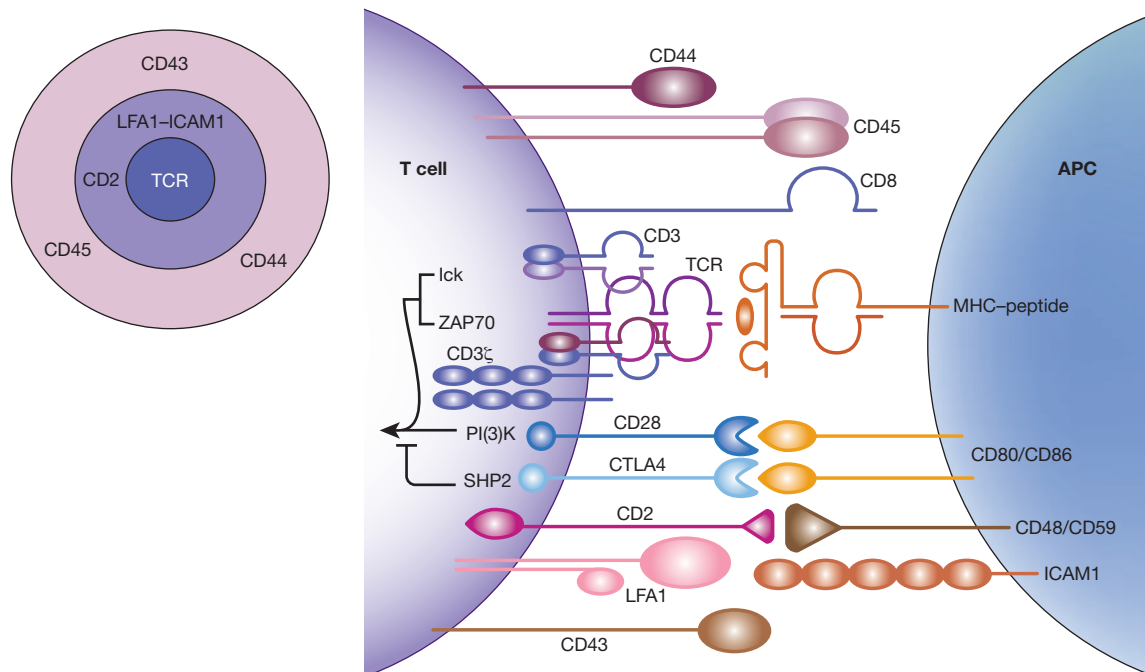


Figure 4 The T-cell synapse. Molecules on a T cell and counterstructures on an antigen-presenting cell (APC) that are known to be involved in T-cell activation are shown. Far left, the face of the synapse with its characteristic distribution of cell-surface molecules, which form a supramolecular activation complex (SMAC). Some of the molecules that are enriched in different regions of the SMAC are shown. The T-cell receptor (TCR) occupies a central region known as the cSMAC; a peripheral ring, the pSMAC, contains adhesion molecules. CTLA4, cytotoxic-T-lymphocyte antigen 4; ICAM1, intercellular adhesion molecule 1; LFA1, lymphocyte-function-associated antigen 1; MHC, major histocompatibility complex; PI(3)K, phosphatidylinositol-3-OH kinase; SHP2, SH2-domain-containing protein tyrosine phosphatase; ZAP70, ζ-chain (TCR)-associated protein kinase, 70 kDa.

events, such as phosphorylation, elicitation of a calcium transient, or alterations in the subcellular localization of key components of the signal transduction machinery⁶⁰. The formation of the immunological synapse begins with integrin-dependent adhesion, closely followed by a redistribution of the TCR, protein kinase C-θ and adhesion molecules^{61,62}. In the APC, the counterstructures for cell-adhesion molecules, the ligand for the TCR and the MHC molecules show similar subcellular redistribution^{42,63,64}. Events that follow include the transcription of diagnostic sets of genes that are turned on, currently measured using genome-wide arrays. For many microbes, the sets of genes expressed under the artificial conditions of the Petri dish differ profoundly from the expression pattern that occurs in the course of a natural infection with that microorganism. T cells maintained *in vitro*, regardless of the exact culture conditions, are also likely to provide at best an approximation of the changes in gene expression that occur when T cells are engaged *in vivo*. Ultimately, T cells acquire functional traits unique to the activated state: the ability to sense the appropriate set of chemotactic cues to reach a site of inflammation, the ability to synthesize and release cytokines that act on T cells themselves and on susceptible target cells within range, or in the case of cytolytic T cells, a licence to kill. T cells are not inert in the absence of a signal delivered by means of the TCR: naive cells perceive chemokine signals and respond to them by leaving the circulation and taking up residence in secondary lymphoid organs⁶⁵. Similarly, not every engagement of the antigen receptor necessarily leads to activation of effector functions: survival signals for T cells include the ligation of TCRs by the MHC products expressed in the host^{66,67}.

Synapses

The concept of the immunological synapse has gained wide currency and its existence is amply supported by observations *in vitro*. Although there is no agreement on their function, supramolecular clusters of diverse surface molecules, as found in the immunological

synapse, must affect the signals transduced by the individual components, even though active signalling molecules rapidly vanish from the synapse⁶⁸ (Fig. 4). The functional distinctions between anergic T cells and their normal counterparts also manifest themselves at the level of synapse formation^{69,70}. In addition, synapses formed by the different types of T cell may depend on the target cell engaged^{68,71,72}, and because of the involvement of non-identical sets of receptors, will also differ between T cells and natural killer cells^{73–75}, even when similar targets are involved. Whatever their role, whether or not such synapses also form *in vivo* is an important question that has yet to be answered. The architecture of the synapses formed by cells deficient in LFA1 (lymphocyte-function-associated antigen 1) is not known, but because LFA1-deficient mice show no defects in positive and negative selection of developing T lymphocytes⁷⁶, at least the signals transduced in the course of T-cell development can be detected and interpreted in an LFA1-independent manner.

During the initial T-cell priming events, membrane proteins at the cell surface of T cells form organized structures, and corresponding counterstructures occur on the APC. Dendritic cells play an active part in the formation of the synapse. Dendritic cells actively polarize their actin cytoskeleton during interaction with a T cell. The rearrangement of the cytoskeleton in dendritic cells is critical for both clustering and T-cell activation⁷⁷. When B cells are used as APCs, no polarization of their cytoskeleton is seen when cognate interaction with T cells occurs, and pharmacological inhibition of rearrangement of the B-cell cytoskeleton has no effect on B-cell/T-cell conjugate formation or on T-cell activation^{78,79}.

The TCR, MHC–peptide, and costimulation and signal transduction molecules are segregated in the central region of the synapse, whereas molecules involved in adhesion localize to the periphery of the synapse. The TCRs first engage MHC products predominantly by TCR–MHC contacts. Subsequent stabilization of these interactions involves a close apposition of the TCR's CDR3 (complementarity

determining region 3) regions with the peptide cargo of the MHC product. In the course of synapse formation, class II MHC molecules accumulate at the B-cell/T-cell interface in a manner that is dependent on antigen and the expression of costimulatory molecules. The latter may stimulate in T cells the surface display of molecules involved in synapse formation. MHC-peptide complexes that would not by themselves activate the T cell contribute to the formation of a stable synapse and so indirectly support T-cell activation. So far, none of the aspects of synapse formation has been recorded by intravital microscopy, a gap in our knowledge that surely must be filled.

The complexity of interactions between T cells and APCs is compounded by the organization of lymphoid tissue. Cells packed closely together nonetheless manage to migrate at considerable speed, and given the modest numbers of naive lymphocytes specific for any given antigen, it would be remarkable that any successful interactions occurred at all were it not for the precise positional cues conferred by chemokines. Migration of activated T cells in tissues is determined by the chemokine directional code⁸⁰. Although the theoretical framework for chemotaxis and alterations in the direction of migration of motile cells have been well worked out and are supported by experiments *in vitro*^{81,82}, migratory behaviour may be different in a whole animal⁸³. The intraparenchymal migration of live T cells has yet to be visualized *in vivo*. The extent to which homologous and heterologous desensitization of the G-protein-coupled chemokine receptors contribute to the navigation of T cells will also benefit from the emerging techniques for intravital imaging.

Naive T cells generally reside in lymphoid tissues, whereas effector and memory T cells migrate preferentially to tissues that are drained by the secondary lymphoid organs where antigen was first encountered⁸⁰. Dendritic cells from Peyer's patches — a collection of lymph tissues located in the mucus-secreting lining of the small intestine — induce the expression in T cells of integrin $\alpha 4\beta 7$ and CCR9, the receptor for gut-associated chemokine TECK/CCL25, to allow effector and memory T cells to migrate to tissue that contains their cognate antigen⁸⁴. Analogous to chemokines, leukotrienes are secreted factors that function as chemoattractants for leukocytes^{85,86}. Leukotriene B₄ is a ligand for the high-affinity receptor BLT1 on myeloid leukocytes, and is expressed by effector CD8 T cells, not by naive or central memory T cells. The binding of leukotriene B₄ to BLT1 on cytotoxic T lymphocytes is now recognized as a non-chemokine pathway for the migration of CD8 effector cells.

ERM proteins and cytoskeletal elements

The migration of T cells in lymph nodes and other tissues requires not only chemotactic cues, but also the modulation of cell adhesion, and the reorganization of components of the extracellular matrix and cytoskeleton. Members of the widely expressed ezrin-radixin-moesin (ERM) family of cytoskeletal proteins act as general cross-linkers between the cortical actin filament network and the plasma membrane^{87,88}. Their involvement in T-cell activation is well understood and operates mainly through their binding to the cytoplasmic regions of integral membrane proteins such as CD43 in T lymphocytes^{89,90}. Their involvement in APC morphology and function during T-cell ligation is almost unexplored.

In epithelial cells, the ligation of adhesion molecules initially establishes cell polarity⁹¹, which is followed by the organization of the cytoskeleton and sorting of proteins to basolateral and apical compartments^{92,93}. Protein clustering at the cortical membranes occurs, in part, through binding to submembrane scaffolding proteins, which include ERM proteins. Genetic studies in *Drosophila* and *Caenorhabditis elegans* show that ERM orthologues polarize to apical and basolateral compartments and are fixed to the cytoskeleton, where they perform structural and regulatory functions.

Patterns of expression of the three most conserved ERMs are similar in a variety of cells and occur at projections such as microvilli, ruffled membranes, uropods or growth cones. When polarized at these sites, ERM proteins maintain an activated conformation that

links cortical membrane proteins to cytoskeletal actin through conserved amino- and carboxy-terminal ERM-association domains (ERMADs)⁹⁴. The N-ERMAD can bind directly to plasma membrane proteins in the case of ICAM1 (intercellular adhesion molecule 1), ICAM2, ICAM3, CD43 and CD44, or indirectly, through adaptors, which provide docking sites for transmembrane proteins (for example, the cystic fibrosis transmembrane conductance regulator, the β_2 -adrenergic receptor and the sodium-hydrogen exchanger NHE3)⁹⁵.

The cognate interaction of T cells and APCs induces cytoskeletal changes in T cells that appear to adapt the shape of the T-cell surface to that of the APC. Such cytoskeletal changes occur through changes in the phosphorylation status of ERM proteins, which disrupt the anchoring of the cortical actin cytoskeleton from the plasma membrane, and so decrease cellular rigidity to promote more efficient T-cell-APC conjugate formation.

ERM proteins are not the only proteins that anchor surface proteins to cytoskeletal elements. For example, in vertebrate cells an extensive membrane-associated spectrin-ankyrin scaffold conjugates membrane and cytosolic proteins to the actin and microtubule cytoskeleton⁹⁶. Transport mediated by spectrin-ankyrin is important because it facilitates the movement of CD45 (and CD3) to the lymphocyte surface. Proximity of CD45 to the TCR is necessary for the activation of T cells, and defects in CD45 underlie human forms of severe combined immunodeficiency^{96,97}. Activation of T cells by the display of class II MHC-peptide complexes by APCs induces reorganization of the T-cell cytoskeleton, which results in the polarization of the microtubule-organizing center (MTOC). Study of MTOC reorientation in real time has been done in Jurkat cell lines that stably express GFP-tubulin (either proficient or deficient in adaptor molecules involved in calcium mobilization, which is essential for T-cell activation). Calcium-dependent events are required for MTOC polarization towards superantigen-laden B cells, although the players involved in calcium-regulated MTOC polarization are not all known as yet.

Interactions of dendritic cells and T cells in the lymph node

The use of intravital microscopy, sometimes combined with real-time two-photon excitation of fluorophores, has been instrumental in migration studies of T cells *in vivo*^{83,98}. The priming of T-cell responses requires conjugate formation between dendritic cells and T cells; the timely dance between dendritic cells and CD4-positive T cells *in vivo* is now a matter of record¹⁵. During the first eight hours after entry from the circulation, T cells and dendritic cells undergo multiple short encounters, which in the presence of antigen last for about six minutes on average (or four minutes in the absence of antigen). Long-lasting, stable interactions of at least 30 minutes' duration occur during the subsequent 12 hours, during which production of the cytokines interleukin-2 and interferon- γ by T cells is initiated¹⁵. After 24 hours, T cells again have short encounters with dendritic cells, now accompanied by the onset of proliferation and rapid migration. Earlier work on intact explanted lymph nodes indicated that interactions between dendritic cells and T cells are mainly stable in duration, lasting for hours^{99,100}. Maintenance of lymphatic flow, an intact circulatory system, and continued innervation of the tissue examined account for the differences between the intravital and *ex vivo* explant approaches.

Technological advances have put immunologists in the enviable position of being able to correlate an enormous amount of functional data obtained *in vitro* for T cells, B cells and dendritic cells with *in vivo* functions. Being able to manipulate the immune system genetically, and to easily produce radiation chimaeras that involve genetically manipulated haematopoietic precursors, provides opportunities not available for any other organ system. Mice can be created in which lymphocytes, or only some lymphocytes, are the sole cells that harbour a mutation of interest, if necessary introduced as a conditional allele. A thorough understanding of T-cell activation *in vivo* is indispensable

able for the development of better vaccines, for our understanding of the regulatory circuits that break down when autoimmunity causes disease, and for a better understanding of our interactions with the microbial world around us. □

doi:10.1038/nature02762

1. von Andrian, U. H. & Mempel, T. R. Homing and cellular traffic in lymph nodes. *Nature Rev. Immunol.* **3**, 867–878 (2003).
2. Itano, A. A. & Jenkins, M. K. Antigen presentation to naive CD4 T cells in the lymph node. *Nature Immunol.* **4**, 733–739 (2003).
3. Barton, G. M. & Medzhitov, R. Control of adaptive immune responses by Toll-like receptors. *Curr. Opin. Immunol.* **14**, 380–383 (2002).
4. Means, R. E., Lang, S. M., Chung, Y. H. & Jung, J. U. Kaposi's sarcoma associated herpesvirus immune evasion strategies. *Front. Biosci.* **7**, e185–e203 (2002).
5. Vossen, M. T., Westerhout, E. M., Soderberg-Naucler, C. & Wiertz, E. J. Viral immune evasion: a masterpiece of evolution. *Immunogenetics* **54**, 527–542 (2002).
6. Seet, B. T. *et al.* Poxviruses and immune evasion. *Annu. Rev. Immunol.* **21**, 377–423 (2003).
7. Tufariello, J. M., Chan, J. & Flynn, J. L. Latent tuberculosis: mechanisms of host and bacillus that contribute to persistent infection. *Lancet Infect. Dis.* **3**, 578–590 (2003).
8. Tortorella, D., Gewurz, B. E., Furman, M. H., Schust, D. J. & Ploegh, H. L. Viral subversion of the immune system. *Annu. Rev. Immunol.* **18**, 861–926 (2000).
9. Davison, A. J. *et al.* The human cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genome. *J. Gen. Virol.* **84**, 17–28 (2003).
10. Borst, E. M., Hahn, G., Koszinowski, U. H. & Messerle, M. Cloning of the human cytomegalovirus (HCMV) genome as an infectious bacterial artificial chromosome in *Escherichia coli*: a new approach for construction of HCMV mutants. *J. Virol.* **73**, 8320–8329 (1999).
11. Itano, A. A. *et al.* Distinct dendritic cell populations sequentially present antigen to CD4 T cells and stimulate different aspects of cell-mediated immunity. *Immunity* **19**, 47–57 (2003).
12. Shortman, K. & Liu, Y. J. Mouse and human dendritic cell subtypes. *Nature Rev. Immunol.* **2**, 151–161 (2002).
13. Princiotta, M. F. *et al.* Quantitating protein synthesis, degradation, and endogenous antigen processing. *Immunity* **18**, 343–354 (2003).
14. Brigl, M. & Brenner, M. B. CD1: antigen presentation and T cell function. *Annu. Rev. Immunol.* **22**, 817–890 (2004).
15. Mempel, T. R., Henrickson, S. E. & Von Andrian, U. H. T-cell priming by dendritic cells in lymph nodes occurs in three distinct phases. *Nature* **427**, 154–159 (2004).
16. Valitutti, S., Muller, S., Cella, M., Padovan, E. & Lanzavecchia, A. Serial triggering of many T-cell receptors by a few peptide MHC complexes. *Nature* **375**, 148–151 (1995).
17. Inaba, K. *et al.* The formation of immunogenic major histocompatibility complex class II–peptide ligands in lysosomal compartments of dendritic cells is regulated by inflammatory stimuli. *J. Exp. Med.* **191**, 927–936 (2000).
18. Pierre, P. *et al.* Developmental regulation of MHC class II transport in mouse dendritic cells. *Nature* **388**, 787–792 (1997).
19. Turley, S. J. *et al.* Transport of peptide MHC class II complexes in developing dendritic cells. *Science* **288**, 522–527 (2000).
20. Lennon-Dumenil, A. M. *et al.* Analysis of protease activity in live antigen-presenting cells shows regulation of the phagosomal proteolytic contents during dendritic cell activation. *J. Exp. Med.* **196**, 529–540 (2002).
21. Trombetta, E. S., Ebersold, M., Garrett, W., Pypaert, M. & Mellman, I. Activation of lysosomal function during dendritic cell maturation. *Science* **299**, 1400–1403 (2003).
22. Aderem, A. How to eat something bigger than your head. *Cell* **110**, 5–8 (2002).
23. Lennon-Dumenil, A. M., Bakker, A. H., Wolf, Bryant, P., Ploegh, H. L. & Lagaudriere-Gesbert, C. A closer look at proteolysis and MHC-class-II-restricted antigen presentation. *Curr. Opin. Immunol.* **14**, 15–21 (2002).
24. Sugita, M., Peters, P. J. & Brenner, M. B. Pathways for lipid antigen presentation by CD1 molecules: nowhere for intracellular pathogens to hide. *Traffic* **1**, 295–300 (2000).
25. Blander, J. M. & Medzhitov, R. Regulation of phagosome maturation by signals from toll-like receptors. *Science* **304**, 1014–1018 (2004).
26. Chow, A., Toomre, D., Garrett, W. & Mellman, I. Dendritic cell maturation triggers retrograde MHC class II transport from lysosomes to the plasma membrane. *Nature* **418**, 988–994 (2002).
27. Boes, M. *et al.* T-cell engagement of dendritic cells rapidly rearranges MHC class II transport. *Nature* **418**, 983–988 (2002).
28. Boes, M. *et al.* T cells induce extended class II MHC compartments in dendritic cells in a toll-like receptor-dependent manner. *J. Immunol.* **171**, 4081–4088 (2003).
29. Bertho, N. *et al.* Requirements for T cell-polarized tubulation of class II⁺ compartments in dendritic cells. *J. Immunol.* **171**, 5689–5696 (2003).
30. Kleijmeer, M. *et al.* Reorganization of multivesicular bodies regulates MHC class II antigen presentation by dendritic cells. *J. Cell Biol.* **155**, 53–63 (2001).
31. Cherukuri, A., Dykstra, M. & Pierce, S. K. Floating the raft hypothesis: lipid rafts play a role in immune cell activation. *Immunity* **14**, 657–660 (2001).
32. Horejsi, V. The roles of membrane microdomains (rafts) in T cell activation. *Immunol. Rev.* **191**, 148–164 (2003).
33. Vogt, A. B., Spindeldreher, S. & Kropshofer, H. Clustering of MHC peptide complexes prior to their engagement in the immunological synapse: lipid raft and tetraspan microdomains. *Immunol. Rev.* **189**, 136–151 (2002).
34. Pierce, S. K. To cluster or not to cluster: FRETting over rafts. *Nature Cell Biol.* **6**, 180–181 (2004).
35. Munro, S. Lipid rafts: elusive or illusive? *Cell* **115**, 377–388 (2003).
36. Glebov, O. O. & Nichols, B. J. Lipid raft proteins have a random distribution during localized activation of the T-cell receptor. *Nature Cell Biol.* **6**, 238–243 (2004).
37. Paladino, S., Sarnataro, D. & Zurzolo, C. Detergent-resistant membrane microdomains and apical sorting of GPI-anchored proteins in polarized epithelial cells. *Int. J. Med. Microbiol.* **291**, 439–445 (2002).
38. Jan, Y. N. & Jan, L. Y. The control of dendrite development. *Neuron* **40**, 229–242 (2003).
39. Tarrant, J. M., Robb, L., van Spriel, A. B. & Wright, M. D. Tetraspanins: molecular organisers of the leukocyte surface. *Trends Immunol.* **24**, 610–617 (2003).
40. Stipp, C. S., Kolesnikova, T. V. & Hemler, M. E. Functional domains in tetraspanin proteins. *Trends Biochem. Sci.* **28**, 106–112 (2003).
41. Kropshofer, H. *et al.* Tetraspan microdomains distinct from lipid rafts enrich select peptide MHC class II complexes. *Nature Immunol.* **3**, 61–68 (2002).
42. Heath, W. R. & Carbone, F. R. Cross-presentation, dendritic cells, tolerance and immunity. *Annu. Rev. Immunol.* **19**, 47–64 (2001).
43. Sigal, L. J., Crotty, S., Andino, R. & Rock, K. L. Cytotoxic T-cell immunity to virus-infected non-haematopoietic cells requires presentation of exogenous antigen. *Nature* **398**, 77–80 (1999).
44. van Endert, P. M. Peptide selection for presentation by HLA class I: a role for the human transporter associated with antigen processing? *Immunol. Res.* **15**, 265–279 (1996).
45. Wiertz, E. J. *et al.* Sec61-mediated transfer of a membrane protein from the endoplasmic reticulum to the proteasome for destruction. *Nature* **384**, 432–438 (1996).
46. Johnson, A. E. & van Waas, M. A. The translocin: a dynamic gateway at the ER membrane. *Annu. Rev. Cell Dev. Biol.* **15**, 799–842 (1999).
47. Ackerman, A. L., Kyritsis, C., Tampe, R. & Cresswell, P. Early phagosomes in dendritic cells form a cellular compartment sufficient for cross-presentation of exogenous antigens. *Proc. Natl Acad. Sci. USA* **100**, 12889–12894 (2003).
48. Gagnon, E. *et al.* Endoplasmic reticulum-mediated phagocytosis is a mechanism of entry into macrophages. *Cell* **110**, 119–131 (2002).
49. Guernonprez, P. *et al.* ER-phagosome fusion defines an MHC class I cross-presentation compartment in dendritic cells. *Nature* **425**, 397–402 (2003).
50. Houde, M. *et al.* Phagosomes are competent organelles for antigen cross-presentation. *Nature* **425**, 402–406 (2003).
51. Lilley, B. N. & Ploegh, H. L. A membrane protein required for dislocation of misfolded proteins from the ER. *Nature* **429**, 834–840 (2004).
52. Binder, R. J., Blachere, N. E. & Srivastava, P. K. Heat shock protein-chaperoned peptides but not free peptides introduced into the cytosol are presented efficiently by major histocompatibility complex I molecules. *J. Biol. Chem.* **276**, 17163–17171 (2001).
53. Wolkers, M. C., Brouwenstijn, N., Bakker, A. H., Toebes, M. & Schumacher, T. N. Antigen bias in T cell cross-priming. *Science* **304**, 1314–1317 (2004).
54. Norbury, C. C. *et al.* CD8⁺ T cell cross-priming via transfer of proteasome substrates. *Science* **304**, 1318–1321 (2004).
55. Ploegh, H. L. Immunology: Nothing 'gainst time's scythe can make defense. *Science* **304**, 1262–1263 (2004).
56. Clark, R. H. *et al.* Adaptor protein 3-dependent microtubule-mediated movement of lytic granules to the immunological synapse. *Nature Immunol.* **4**, 1111–1120 (2003).
57. Kuhne, M. R. *et al.* Linker for activation of T cells, ζ -associated protein-70, and Src homology 2 domain-containing leukocyte protein-76 are required for TCR-induced microtubule-organizing center polarization. *J. Immunol.* **171**, 860–866 (2003).
58. Lewis, R. S. Calcium signaling mechanisms in T lymphocytes. *Annu. Rev. Immunol.* **19**, 497–521 (2001).
59. Panyi, G. *et al.* Kv1.3 potassium channels are localized in the immunological synapse formed between cytotoxic and target cells. *Proc. Natl Acad. Sci. USA* **101**, 1285–1290 (2004).
60. Hogan, P. G., Chen, L., Nardone, J. & Rao, A. Transcriptional regulation by calcium, calcineurin, and NFAT. *Genes Dev.* **17**, 2205–2232 (2003).
61. Dustin, M. L. The immunological synapse. *Arthritis Res.* **4** (suppl. 3), S119–S125 (2002).
62. Krogsgaard, M., Huppa, J. B., Purbhoo, M. A. & Davis, M. M. Linking molecular and cellular events in T-cell activation and synapse formation. *Semin. Immunol.* **15**, 307–315 (2003).
63. Anderson, H. A., Hiltbold, E. M. & Roche, P. A. Concentration of MHC class II molecules in lipid rafts facilitates antigen presentation. *Nature Immunol.* **1**, 156–162 (2000).
64. Wetzel, S. A., McKeithan, T. W. & Parker, D. C. Live-cell dynamics and the role of costimulation in immunological synapse formation. *J. Immunol.* **169**, 6092–6101 (2002).
65. Weninger, W. & von Andrian, U. H. Chemokine regulation of naive T cell traffic in health and disease. *Semin. Immunol.* **15**, 257–270 (2003).
66. Ernst, B., Lee, D. S., Chang, J. M., Sprent, J. & Surh, C. D. The peptide ligands mediating positive selection in the thymus control T cell survival and homeostatic proliferation in the periphery. *Immunity* **11**, 173–181 (1999).
67. Kieper, W. C., Burghardt, J. T. & Surh, C. D. A role for TCR affinity in regulating naive T cell homeostasis. *J. Immunol.* **172**, 40–44 (2004).
68. Lee, K. H. *et al.* The immunological synapse balances T cell receptor signaling and degradation. *Science* **302**, 1218–1222 (2003).
69. Huppa, J. B., Gleimer, M., Sumen, C. & Davis, M. M. Continuous T cell receptor signaling required for synapse maintenance and full effector potential. *Nature Immunol.* **4**, 749–755 (2003).
70. Ehrlich, L. I., Ebert, P. J., Krummel, M. F., Weiss, A. & Davis, M. M. Dynamics of p56lck translocation to the T cell immunological synapse following agonist and antagonist stimulation. *Immunity* **17**, 809–822 (2002).
71. Somersalo, K. *et al.* Cytotoxic T lymphocytes form an antigen-independent ring junction. *J. Clin. Invest.* **113**, 49–57 (2004).
72. Favier, B. *et al.* Uncoupling between immunological synapse formation and functional outcome in human $\gamma\delta$ T lymphocytes. *J. Immunol.* **171**, 5027–5033 (2003).
73. Orange, J. S. *et al.* The mature activating natural killer cell immunologic synapse is formed in distinct stages. *Proc. Natl Acad. Sci. USA* **100**, 14151–14156 (2003).
74. Davis, D. M. Assembly of the immunological synapse for T cells and NK cells. *Trends Immunol.* **23**, 356–363 (2002).
75. McCann, F. E. *et al.* The size of the synaptic cleft and distinct distributions of filamentous actin, ezrin, CD43, and CD45 at activating and inhibitory human NK cell immune synapses. *J. Immunol.* **170**, 2862–2870 (2003).
76. Shier, P. *et al.* Impaired immune responses toward alloantigens and tumor cells but normal thymic selection in mice deficient in the $\beta 2$ integrin leukocyte function-associated antigen-1. *J. Immunol.* **157**, 5375–5386 (1996).
77. Al-Alwan, M. M. *et al.* Cutting edge: dendritic cell actin cytoskeletal polarization during immunological synapse formation is highly antigen-dependent. *J. Immunol.* **171**, 4479–4483 (2003).

78. Kupfer, A. & Singer, S. J. The specific interaction of helper T cells and antigen-presenting B cells. IV. Membrane and cytoskeletal reorganizations in the bound T cell as a function of antigen dose. *J. Exp. Med.* **170**, 1697–1713 (1989).
79. Wulfig, C., Sjaastad, M. D. & Davis, M. M. Visualizing the dynamics of T cell activation: intracellular adhesion molecule 1 migrates rapidly to the T cell/B cell interface and acts to sustain calcium levels. *Proc. Natl Acad. Sci. USA* **95**, 6302–6307 (1998).
80. Sallusto, F. & Lanzavecchia, A. Understanding dendritic cell and T-lymphocyte traffic through the analysis of chemokine receptor expression. *Immunol. Rev.* **177**, 134–140 (2000).
81. Foxman, E. F., Kunkel, E. J. & Butcher, E. C. Integrating conflicting chemotactic signals. The role of memory in leukocyte navigation. *J. Cell Biol.* **147**, 577–588 (1999).
82. Wong, M. M. & Fish, E. N. Chemokines: attractive mediators of the immune response. *Semin. Immunol.* **15**, 5–14 (2003).
83. Miller, M. J., Wei, S. H., Cahalan, M. D. & Parker, I. Autonomous T cell trafficking examined *in vivo* with intravital two-photon microscopy. *Proc. Natl Acad. Sci. USA* **100**, 2604–2609 (2003).
84. Mora, J. R. *et al.* Selective imprinting of gut-homing T cells by Peyer's patch dendritic cells. *Nature* **424**, 88–93 (2003).
85. Randolph, G. J. Dendritic cell migration to lymph nodes: cytokines, chemokines, and lipid mediators. *Semin. Immunol.* **13**, 267–274 (2001).
86. Goodarzi, K., Goodarzi, M., Tager, A. M., Luster, A. D. & von Andrian, U. H. Leukotriene B₄ and BLT1 control cytotoxic effector T cell recruitment to inflamed tissues. *Nature Immunol.* **4**, 965–973 (2003).
87. Tsukita, S. & Yonemura, S. Cortical actin organization: lessons from ERM (ezrin/radixin/moesin) proteins. *J. Biol. Chem.* **274**, 34507–34510 (1999).
88. Faure, S. *et al.* ERM proteins regulate cytoskeleton relaxation promoting T cell–APC conjugation. *Nature Immunol.* **5**, 272–279 (2004).
89. Delon, J., Kaibuchi, K. & Germain, R. N. Exclusion of CD43 from the immunological synapse is mediated by phosphorylation-regulated relocation of the cytoskeletal adaptor moesin. *Immunity* **15**, 691–701 (2001).
90. Roumier, A. *et al.* The membrane-microfilament linker ezrin is involved in the formation of the immunological synapse and in T cell activation. *Immunity* **15**, 715–728 (2001).
91. Ojakian, G. K., Ratcliffe, D. R. & Schwimmer, R. Integrin regulation of cell–cell adhesion during epithelial tubule formation. *J. Cell Sci.* **114**, 941–952 (2001).
92. Speck, O., Hughes, S. C., Noren, N. K., Kulikaukas, R. M. & Fehon, R. G. Moesin functions antagonistically to the Rho pathway to maintain epithelial integrity. *Nature* **421**, 83–87 (2003).
93. Bretscher, A., Edwards, K. & Fehon, R. G. ERM proteins and merlin: integrators at the cell cortex. *Nature Rev. Mol. Cell Biol.* **3**, 586–599 (2002).
94. Bretscher, A., Chambers, D., Nguyen, R. & Reczek, D. ERM–Merlin and EBP50 protein families in plasma membrane organization and function. *Annu. Rev. Cell Dev. Biol.* **16**, 113–143 (2000).
95. Short, D. B. *et al.* An apical PDZ protein anchors the cystic fibrosis transmembrane conductance regulator to the cytoskeleton. *J. Biol. Chem.* **273**, 19797–19801 (1998).
96. Pradhan, D. & Morrow, J. The spectrin–ankyrin skeleton controls CD45 surface display and interleukin-2 production. *Immunity* **17**, 303–315 (2002).
97. Buckley, R. H. Advances in the understanding and treatment of human severe combined immunodeficiency. *Immunol. Res.* **22**, 237–251 (2000).
98. Stein, J. V. *et al.* The CC chemokine thymus-derived chemotactic agent 4 (TCA-4, secondary lymphoid tissue chemokine, 6Ckine, exodus-2) triggers lymphocyte function-associated antigen 1-mediated arrest of rolling T lymphocytes in peripheral lymph node high endothelial venules. *J. Exp. Med.* **191**, 61–76 (2000).
99. Bousso, P. & Robey, E. Dynamics of CD8⁺ T cell priming by dendritic cells in intact lymph nodes. *Nature Immunol.* **4**, 579–585 (2003).
100. Stoll, S., Delon, J., Brotz, T. M. & Germain, R. N. Dynamic imaging of T cell–dendritic cell interactions in lymph nodes. *Science* **296**, 1873–1876 (2002).

Acknowledgements We thank J. Cerny, R. Massol and T. Kirchhausen for help with figure 1.

Competing interests statement The authors declare that they have no competing financial interests.

Contacts

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:
Nevin Bayourni (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)
Andy Douglas (4975)
Frank Phelan (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)
Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Production Manager:

Billie Franklin
To send materials use London
address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Stefan Hales

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan, Odo Wulffen
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mail: p.phelan@nature.com
o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@nature.com

US Sales Manager:

Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatomachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami
e-mail: rasami@nature.com

naturejobs

Brought to book

If a book makes it onto the bestsellers list, the publishers can be confident that they have struck a chord. For foundations, which generate a constant stream of white papers and reports, success is best quantified in terms of Internet downloads. In this vein, a book published earlier this year by the Howard Hughes Medical Institute (HHMI) in Chevy Chase, Maryland, has earned a 'greatest clicks' award for career advice.

The popularity of the institute's *Making the Right Moves: A Practical Guide to Scientific Management for Postdocs and New Faculty* says something about how much help young scientists feel they need as they face one of their biggest career challenges — moving from a postdoc to becoming an independent researcher (www.hhmi.org/grants/office/graduate/labmanagement.html). The hunger for information is clear from the number of hits the HHMI initially received: in February and March, about 10,000 copies of the book were downloaded each month.

The need for this book became apparent a few years ago, when the HHMI canvassed its new investigators to find out what they saw as their greatest skill gap. Focus groups had revealed that there was a lag time of several months before grantees received their funding and were finally up and running. Many said that a lack of lab management and grant-writing skills were to blame, so, working with the Burroughs Wellcome Fund, the HHMI engineered a course to address the issue. As the course's popularity grew, a book based on the teaching materials was put together.

The next step may be to modify the training advice to meet the requirements of different countries' career paths. To the HHMI's credit, it has built the electronic version of the book in a modular fashion, and is encouraging people to adapt it to their own needs. This 'shareware' approach to careers advice means that the HHMI is likely to have an electronic hit on its hands for many months.

Paul Smaglik
Naturejobs editor



Contents

REGIONS

Montreal breathes life
back into its science p274

CAREER VIEW

Recruiters & Industry

Building biostatistics

Graduate Journal

Learning to supervise

Movers

Simon Bright p276

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Coming back to life Montreal

The phosphate backbones of a massive DNA helix curl their way around a concrete column, reaching towards the ceiling of Thomas Hudson's office. The art installation may be a conversation piece, but it also indicates how important this little molecule is to Hudson and to Montreal. In 2003, Hudson opened the McGill University and Genome Québec Innovation Centre, a glass and limestone building nestled between the university campus and the city's famous Mount Royal Park. Its more than 100 employees, from graduate students to senior scientists, plan to put Montreal at the hub of the world's genomics and proteomics map.

Montreal has always held the life sciences close to its heart — multinational pharmaceutical companies such as Abbott Laboratories, Novartis and Merck Frosst line the Trans-Canada Highway that runs through the middle of the island. But recently, government and private sector investment has helped establish academic research facilities dedicated to cancer biology, genomics and proteomics in the heart of the city. As well as foreign investment, the boost could bring some Canadian-trained brains back home.



Thomas Hudson: bringing expertise together.

INTELLECTUAL CAPITAL

In 2002, Montreal had more university research investment, at Can\$802 million (US\$587 million), than any other Canadian city. Last year, two of its universities — Montreal and McGill — ranked among the six best-funded Canadian research universities. According to Robert Lacroix, rector of the University of Montreal, the grey-matter concentration here is second only to that of Boston in North America: Montreal has 4.17 students for every 100 residents.

Although Montreal's universities are now at the



T. BOGNAR/CORBIS

heart of its research community, it wasn't always so. Students graduated with exceptional educations, but went away to find jobs or postgraduate positions. Hudson says that most of the changes began about five years ago. "There was a realization that we can do things here, we can develop and we can do the best," says Hudson. "Before that people always had the attitude: 'Let it happen in the United States.'"

Backing that change of view, federal and provincial funding for research infrastructure has soared. The McGill centre has received Can\$40 million from the Canadian Foundation for Innovation (CFI), an independent body set up with a Can\$3.65-billion operating budget to strengthen the infrastructure of Canadian research. Another Can\$52 million has come from not-for-profit provincial corporations, including

Nano state

Aerospace, telecommunications and pulp and paper are Quebec's core industries, and nanotechnology is increasingly important to all of them. So it is no surprise that Montreal has a strong nanoscience presence. "Montreal is going to be one of the major hubs of nanotech development in Canada," says Robert Sing, administrative director of NanoQuebec, the province's university network in nanoscience and nanotechnology.

Montreal's nanotech research community has taken a broad approach to strengthening the field. The Quebec college system recently introduced a training programme that would produce graduates with technical training in nanomaterials, nanobiotechnology and nanofabrication. At university level, there has been significant infrastructure funding from the Canadian Foundation for Innovation, NanoQuebec and provincial and federal governments. One product of this increased funding is the

J.-Armand Bombardier pavilion, which will house an interdisciplinary group of more than 700 chemists, physicists and engineers dedicated to nanoscience and nanotechnology research, with more than Can\$200 million (US\$146 million) of equipment.

But Neil Gordon, president of the Canadian NanoBusiness Alliance, says that there has been a failure to capitalize on the commercialization of knowledge. "We are creating excellent opportunities for students and government research, but we're not creating jobs," he says. He is pushing for Can\$100-million government investment at the commercial and industrial level.

Although this would be a significant leap, it isn't unthinkable. For the first time in 30 years, the prime minister now has a national science adviser. Formerly head of the National Research Council, science adviser Arthur Carty will be expected, among other duties, to foster commercialization.

H.H.



Valorisation-Recherche Québec, which helps finance and market university research, and Genome Québec. The money has funded robotics, high-throughput sequencing systems and shelves of machines for amplifying DNA using the polymerase chain reaction.

In return, the innovation centre has been good for the economy. Within 21 months of opening, it had generated more than Can\$4 million in revenue from its four core services: sequencing, genotyping, DNA microchips and proteomics. In May 2004, its status as one of the few centres in the world to have integrated the Illumina genotyping system won it a Can\$1.2 million service agreement with GenoMed, a genomics company based in St Louis, Missouri.

The McGill information centre occupies one corner of a high-tech courtyard. Close by are the chemical engineering department's MH Wong building, the Ernest Rutherford physics building (the New Zealand Nobel laureate spent seven years teaching at McGill) and a new information technology department. "To be able to ask large-scale questions with genomics and proteomics, you need to build large-scale infrastructure and you need mathematicians, chemists and biologists," says Hudson. "We planned this environment. We wanted to bring cores of expertise together in one place."

Across Mount Royal Park, the University of Montreal is constructing Canada's first systems-biology institute. The Institute for Research in Immunovirology and Cancer (IRIC) aims to let researchers from different fields interact seamlessly. Its building is linked to the university's bioinformatics building and next to the J.-Armand Bombardier Pavilion: a massive new nanosciences institute (see box, opposite).

The Can\$100-million IRIC building, funded largely by CFI grants, is due to open in September with 200

Bright future: Montreal's city hall is at the heart of a community whose commitment to innovative research is calling scientific exiles home.

students, researchers and technicians in biochemistry, biology, genetics, structural biology and immunology. "We want to be able to continue recruiting according to the needs and directions of research," says director Pierre Chartrand. "Closer to 2008 there will be about 40 research groups working at the institute."

Chartrand intends to make the IRIC one of the strongest institutes in Canada. Its high-tech environment will include platforms for proteomics, genetic manipulation and cellular and structural biology, not to mention an entire spectrum of organisms from bacteria to mice — including a 100,000-mouse transgenic facility.

LURE OF THE LAB

"We've planned it so that from day one, the researchers will walk into the lab and be able to start working — it will be fully equipped," says Chartrand. He believes that the impressive infrastructure and resources will attract new people to Montreal.

In the past, Canada has lost some of its best minds to the United States and Europe, but the investment may change that. Hudson, formerly assistant director of the Whitehead Institute/Massachusetts Institute of Technology Genome Center, was lured home to Canada by new opportunities. Another exile, Canadian cancer biologist Katherine Borden, is returning from New York's Mount Sinai School of Medicine to join the IRIC in September. "I feel many things for Canadian science have changed," she says.

One change is the federal government's dedication to promoting research excellence by developing well-funded research professorships across the country. In 2000, the government earmarked Can\$900 million to establish 2,000 Canada Research Chairs. Borden will be one of seven IRIC staff to hold one of these, along with researchers from the United States and Europe.

At Montreal International, a non-profit body set up to attract investment and international organizations, vice-president for life sciences Michel Leblanc is trying to capitalize on these new opportunities. "The life-sciences sector in Montreal is very well integrated," he says, meaning that there is an unbroken chain from basic to clinical research, with a wide variety of work opportunities along the way.

Strong ties between industry and academia bring international drug companies to the region, where, says Leblanc, they benefit from the province's system of post-secondary colleges of general and vocational education. "Quebec's pre-university training produces strong technicians," says Leblanc, adding that this makes Montreal a popular destination for European companies. "They'll find the manpower they need."

Montreal's provincial charm is often overshadowed by Toronto's flash. But in reality, Toronto and Montreal are nose-to-nose in the life-sciences and biotech sectors. "Toronto has a lot more generic jobs in the drug industry, whereas Montreal is very active in new innovative pharmaceutical companies and in-house research," says Leblanc. "And the biotech sector is larger in Montreal, in terms of jobs, venture capital and number of companies."

Hannah Hoag is a science writer based in Montreal, Canada.

GRADUATE JOURNAL

Learning to supervise

Andre is an undergraduate I currently supervise. He is working in my lab for six months before he graduates and begins his PhD thesis. Recently, he reassured me by saying that I do a good job as a supervisor — I think that learning how to supervise is a responsibility and skill one should learn early on in one's career.

Although some people feel that training young scientists is a nuisance or a waste of time and energy, I see a challenge in contributing to the development of a potentially excellent scientist. Also, one can benefit from taking on this role. Over the past few years a couple of our students have earned publications — which ultimately benefits the lab.

Supervision needs good communication and an open-minded and fair approach. My experience has taught me that science is basically knowledge management, so you must talk with your whole team regularly, to point out recent developments and identify threats and opportunities to the project.

Such communication isn't always the norm — perhaps because some academics prefer to work alone and others use a 'hands-off' approach. So I have learned the importance of compatibility — whether you will be supervising or being supervised. ■

Philipp Angerer is a second-year PhD student in biotechnology at the Swiss Federal Institute of Technology (ETH) in Zurich, Switzerland.

RECRUITERS & INDUSTRY

Building biostatistics

As drug companies incorporate pharmacogenetics into their development mix, they are finding themselves short of key skills. The genome-wide analysis at the heart of pharmacogenetics needs scientists comfortable at the intersection of statistics, mathematics, informatics, genetics and biology. They also need to develop and apply large-scale analysis, and ensure that the results are meaningful and relevant.

Essential techniques in this field include text mining, systems biology, advanced multivariate statistics of gene–gene interaction, genetic epidemiology and statistical genetics, statistical data mining for microarray analysis, genome mining for gene discovery, and mathematical modelling of complex biological systems. This set of skills is

difficult to find in a single person, and interdisciplinary training programmes are only beginning to emerge. Also, in the past, statistics and applied mathematics have become separated; biology needs both.

Even this demanding set of technical skills is not enough. The ideal pharmacogenetics researcher needs excellent communication skills to talk to physicians about phenotypes, to discuss algorithms with mathematicians and explore data analysis and integration with biostatisticians and bioinformaticians.

As academia hasn't been producing these kinds of people, industry may have to make its own. That is the strategy the Serono Genetics Institute (SGI) in Evry, France, is taking — but with a twist.

The SGI needs to boost its 100-scientist staff by 10% this year, particularly in biostatistics for complex-

trait genetics and in bioinformatics, to support its pharmacogenetics research.

How to accomplish this? First, the SGI will use as many recruiting tools as possible to find senior and experienced scientists. It is also looking for young scientists who excel in a skill the company needs and have set up their research or career paths early, or shown curiosity about interdisciplinary research. And to meet the target of ten more experts in biostatistics and bioinformatics this year — and probably more in the future — the SGI will train its own future scientists by offering PhDs and postdocs the possibility of tenure.

So, young scientists should have the courage to build up an unconventional and interdisciplinary career that could lead them into pharmacogenetics. ■

Anne Gimalac is communication manager at Serono Genetics Institute, Evry, France.

MOVERS

Simon Bright, director, Warwick HRI, University of Warwick, UK



Simon Bright's career in plant science mirrors the evolution of the field, as its intellectual vanguard shifted from botany to biochemistry, and from molecular biology to biotechnology, and its professional frontiers shifted from academia to industry and back again (see *Naturejobs* 4–5; 10 October 2002).

His career didn't start out in plant science. A childhood curiosity about chemistry led Bright to an undergraduate

education at the University of Cambridge and a PhD under Don Northcote. His mentor's attitude of "go out and make it happen yourself" has served as a major source of inspiration, says Bright, as has keeping in touch with 150 or so other scientists who did PhDs with Northcote.

Bright went straight into a staff research job after his PhD. "I skipped the postdoc completely, which was very lucky," he says. Rothamsted Research, near London, had just created three new positions to bring new skills into its mix of agricultural research.

He realized that biochemistry could help him understand problems in plant metabolism, but that only molecular biology and biotechnology would enable him to do something with the knowledge, such as change the nutritional composition of plants. By the early 1980s "it was clear that the next thing I had to do was molecular biotech," he says.

He took a one-year sabbatical at

Arco PCRI in the San Francisco Bay Area, where the "extraordinary hub of entrepreneurship and scientific creativity" was so beguiling that he almost didn't return to England. But he decided there were opportunities at home, at a time when the first genetically modified crop plants were being created. He found it "very exciting" to be doing more than just cloning genes, and considers his contributions to the first genetically modified product in the UK — a tomato paste — a highlight in his career.

In industry, Bright learned valuable skills in project management that he will bring back to the academic sector. He is making the switch in part because life-sciences companies are now more interested in development than in basic research. He is convinced that there are discoveries "below the radar" of product-oriented plant-science companies that he calls "the vegetable equivalent of orphan drugs". ■

CV

2002–04: Head of technology interaction, Syngenta, Jealotts Hill, UK

2000–02: Head of European genomics,

Syngenta (cereal genomics/bioinformatics), Norwich, UK

1996–2000: Technology interaction manager, Novartis, Jealotts Hill, UK

1987–96: Plant biotechnology manager, ICI Seeds/Zeneca Seeds/Zeneca Plant Science, Jealotts Hill, UK

1974–86: Research scientist/principal scientist, Rothamsted Research, UK